Dashboard de seguimiento integral al estado de salud de los datos de una organización

Autores
Constanza Rodríguez Zabala
Nelson Ignacio Moreno Arias

Director
Elio Higinio Cables, Ph.D

Universidad Antonio Nariño
Facultad de Ingeniería de Sistemas
Especialización en Gobierno de Datos
Bogotá D.C
2022

Tabla de contenido

Resumen 5

Palabras clave 5

Abstract 6

Keywords 6

1. Introducción 1

2. Descripción y formulación del problema 3

2.1 Objetivo General 5

2.2 Objetivos Específicos 5

3. Marco referencial 6

3.1 Marco teórico 6

3.2 Estado del Arte 20

Impacto 23

Componente de Innovación 23

4.Metodología 24

5. Desarrollo de la propuesta 31

6.Conclusiones 58

7.Referencias 59

Índice de Figuras

Ilustración 1 Pasos clave para Evaluar la Calidad de los Datos. Fuente: Elaboración Propia adaptado del DAMA DMBOK2 25
Ilustración 2: Flujo procesamiento de datos. Fuente propia 38
Ilustración 3: Conteo registros cargados. Fuente propia 40
Ilustración 4: set de datos cargado. Cuente propia 40
Ilustración 5: Acceso a los datos 42
Ilustración 6: Configuración del set de datos en Talend Open Studio 43
Ilustración 7: Set de datos cargado en MySQL 43
Ilustración 8: Creación del proyecto para evaluar la calidad de datos 44
Ilustración 9: Configuración de reglas para evaluar la calidad de los datos ... 45
Ilustración 10: Configuración de reglas para evaluar la calidad de los datos a través de expresiones regulares..... 46
Ilustración 11: Configuración de reglas para evaluar la calidad de los datos a través de expresiones regulares..... 46
Ilustración 12: Válida que no exista textos en mayúscula 46
Ilustración 13: Valida la Conformidad Email 47
Ilustración 14: Conformidad municipio. 47
Ilustración 15: Validez Fechas 48
Ilustración 16: Valida Caracteres especiales..... 48
Ilustración 17: Columnas perfiladas. 49
Ilustración 18: Columnas perfiladas. 49
Ilustración 19: Columnas perfiladas. 49
Ilustración 20: Reglas para evaluar la calidad de los datos precargadas en Talend Open Studio 50

Ilustración 21: Reglas diseñadas para evaluar la calidad de los datos en Talend Open Studio 50
Ilustración 23: Tipo de Documento 51
Ilustración 24: Número de Documento 52
Ilustración 25: Fecha de nacimiento 52
Ilustración 26: Análisis de causa raíz de problemas de datos en la organización 53
Ilustración 27: Resultado integral. Fuente propia 56

Índice de tablas

Tabla 1: medición frecuencia relativa atributos y dimensiones 3
Tabla 2: diccionario de datos. Fuente Propia 34
Tabla 3: tipos de datos set. Fuente propia 34
Tabla 4 Impacto para el negocio. Fuente: Adaptado del material de DAMA Colombia 37
Tabla 5: Elementos de Datos y Dimensiones. Fuente propia 41
Tabla 6: Calculo peso dimensiones. Fuente propia 54
Tabla 7: ideal de referencia. Fuente propia 55
Tabla 8: Intervalos o rangos. Fuente propia 55
Tabla 9: Matriz de decisión. Fuente propia 55
Tabla 10: Matriz normalizada. Fuente propia 55
Tabla 11: Resultados método RIM. Fuente propia 56

Resumen

La Calidad de los datos dentro de las organizaciones es uno de los temas más importantes a gestionar, ya que sin datos de alta calidad las diferentes capas de la organización se ven afectadas a nivel estratégico, táctico y operativo, esto se refleja en una mala gestión de clientes, reprocesos, sobrecostos, deterioro de la imagen corporativa.

En el presente documento se quiere presentar cómo a través de la aplicación de buenas prácticas las organizaciones pueden realizar un seguimiento integral al estado de salud de los datos, es así como se ha tenido en cuenta lo descrito por la "Guía de Fundamentos de Gestión de Datos" de la Asociación de Gestión de Datos DAMA en su versión 2[1] específicamente en el capítulo de calidad de datos, donde indica que para la mejora continua de calidad de los datos es clave que las organizaciones consideren el ciclo iterativo Deming[2] que significa "Planear, Hacer, Verificar y Actuar".

Es así, como para este ejercicio el enfoque está orientado en la etapa de Evaluación de la Calidad de los Datos, donde se implementa el paso a paso para Evaluar la Calidad de los Datos y entregar a las organizaciones una herramienta de seguimiento que les permita cuantificar la calidad de los datos de forma integral utilizando estándares de datos como dimensiones, métodos a una muestra de elementos de datos (atributos) priorizados.

Palabras clave

Calidad de los datos, Evaluación de la calidad, tomar decisiones, Método RIM.

[1]

[2] Ciclo Deming: Modelo de resolución de problemas conocido como "Planear, Hacer, Revisar y Actuar".

Abstract

Data Quality within Organizations is one of the most important issues to manage, since without high quality data the different layers of the organization are affected at a strategic, tactical and operational level, this is reflected in poor management of clients, reprocesses, cost overruns, deterioration of the corporate image.

In this document we want to present how, through the application of good practices, organizations can carry out a comprehensive monitoring of the state of health of the data, this is how what has been described in the Data Management Fundamentals Guide has been taken into account. of the Data Management Association DAMA in its version 2 specifically in the data quality chapter, where it indicates that for the continuous improvement of data quality it is key that organizations consider the Deming iterative cycle that means "Plan, Do, Do". Check and Act".

Thus, for this exercise, the focus is oriented on the Data Quality Assessment stage, where the step-by-step process for Assessing Data Quality is implemented and delivery to organizations of a monitoring tool that allows them to quantify comprehensive data quality using data standards such as dimensions, methods to a sample of prioritized data elements (attributes).

Keywords

Data quality, Quality assessment, decision making, RIM Method.

1. Introducción

Los datos son el eje central de las organizaciones de todos los sectores económicos e industrias y estas organizaciones tienen algo en común, buscan tener confianza en los datos para tomar decisiones más rápido. No prestar atención a los problemas de la calidad de los datos como, por ejemplo, datos duplicados, datos incompletos, datos contradictorios, repercute negativamente en el logro de las estrategias de negocio y esto les cuesta una gran suma de dinero en el tiempo.

En Colombia cada vez son más las organizaciones que invierten en iniciativas de calidad de datos, más allá del análisis porque desean convertir sus datos en accionables, es decir que puedan tomar la mejor decisión en base a los datos para acelerar los procesos de transformación digital, impulsar nuevos productos, mejorar la experiencia con el cliente, agilizar los procesos de negocio y atraer nuevas ganancias.

Por lo tanto, la calidad es actualmente uno de los activos más importantes dentro de una organización teniendo en cuenta el valor que genera en su utilización para temas como la toma de decisiones y casos de uso, lo anterior apoyados en algunos frameworks que dan algunos lineamientos para temas de calidad.

La Guía de Fundamentos de Gestión de Datos DAMA menciona el marco Strong-Wang(1996) el cual menciona que existen (15) quince dimensiones de calidad de datos que se agrupan en cuatro categorías de calidad, las cuales se aplican a un conjunto de datos definido de acuerdo con la característica y necesidad de cada uno de los atributos, sin embargo, los marcos de referencia definen las buenas prácticas para evaluar la calidad de datos pero no mencionan el cómo se aplican para evaluar la calidad de los datos.

Por lo anterior, se hace necesario entregar el paso a paso para evaluar la calidad y entregar un dashboard que permita hacer seguimiento integral al estado de salud de los datos.

2. Descripción y formulación del problema

Las organizaciones actualmente realizan control de la calidad de los datos a través de la evaluaciones periódicas a los datos, sin embargo obtienen un resultado descriptivo el cual consiste en describir los problemas de calidad presentes en los elementos de datos críticos seleccionados, esto se realiza entregando una tabla que detalla los elementos de datos críticos que fueron objeto de evaluación y las dimensiones de calidad aplicadas a partir de las necesidades expuestas por el negocio donde el resultado es tabulado en forma de medición de frecuencia relativa que básicamente consiste en tomar el número de datos con problemas de calidad sobre el número total de registros analizados, pero esto no entrega a las organizaciones una calificación integral por cada uno de los atributos de acuerdo con las dimensiones aplicadas; estos problemas identificados en los datos impactan la agilidad de los procesos de negocio y así el núcleo de la estrategia de la compañía, los que dificulta analizar la causa raíz de los problemas y plantear las oportunidades de mejora.

En la siguiente tabla se presenta un ejemplo de cómo se realiza actualmente la evaluación de la calidad por elemento de dato y dimensión, donde no se evidencia de manera integral como está la calidad de datos cada elemento.

Dimensión	Completitud			Conformidad			Consistencia			Duplicidad			validez		
Elementos de Datos	Reg Evaluados	Reg con Problemas	Frecuencia Relativa	Reg Evaluados	Reg con Problemas	Frecuencia Relativa	Reg Evaluados	Reg con Problemas	Frecuencia Relativa	Reg Evaluados	Reg con Problemas	Frecuencia Relativa	Reg Evaluados	Reg con Problemas	Frecuencia Relativa
Ciudad	100000	10.000	10%	100000	-		100000	13.220	13%	100000	13.220	13%	100000	-	
Departamento	100000	10.000	10%	100000	-		100000	74.000	74%	100000	74.000	74%	100000	-	
Dirección	100000	22.700	23%	100000	-		100000	41.800	42%	100000	25.000	25%	100000	-	
Email	100000	20.000	20%	100000	23.200	23%	100000	83.170	83%	100000	10.140	10%	100000	-	
Fecha Expedición	100000	10.000	10%	100000	-		100000	-		100000	81.400	81%	100000	20.200	20%
Fecha Nacimiento	100000	-		100000	-		100000	52.600	53%	100000	11.000	11%	100000	58.600	59%
Numero Documento	100000	10.950	11%	100000	-		100000	-		100000	10.950	11%	100000	-	
Primer Apellido	100000	52.600	53%	100000	10.000	10%	100000	52.800	53%	100000	17.700	18%	100000	-	
Primer Nombre	100000	10.000	10%	100000	43.000	43%	100000	83.260	83%	100000	18.900	19%	100000	-	
Telefono1	100000	22.000	22%	100000	22.000	22%	100000	10.000	10%	100000	44.000	44%	100000	22.000	22%
Telefono2	100000	10.000	10%	100000	10.000	10%	100000	10.000	10%	100000	2.000	2%	100000	2.000	2%
Tipo Documento	100000	83.000	83%	100000	-		100000	10.000	10%	100000	-		100000	-	

Tabla 1: medición frecuencia relativa atributos y dimensiones

Teniendo en cuenta lo anterior, y considerando el reto de las organizaciones en convertir los datos en un activo de negocio que permita el logro de la estrategia

corporativa, y estos se encuentran con problemáticas, en diferentes estructuras, surge la necesidad de plantearnos la siguiente pregunta:

¿Cómo realizar el seguimiento integral al estado de salud de los datos de una organización a partir de la identificación de los elementos de datos críticos (atributos) y las dimensiones de calidad de datos aplicables a estos?

2.1 Objetivo General

Elaborar un dashboard que permita el seguimiento integral al estado de salud de los datos de las organizaciones.

2.2 Objetivos Específicos

- Describir los métodos fundamentales existentes para evaluar la calidad de los datos.
- Caracterizar los elementos de datos (atributos) de contactabilidad que son relevantes para las diferentes organizaciones en el área comercial.
- Valorar el conjunto de herramientas que existen en el mercado para el tratamiento de la calidad de datos.
- Implementar el método RIM (método del ideal de referencia) a las dimensiones y elementos de datos priorizados para obtener la calificación de la Calidad de datos de la organización de manera integral.

3. Marco referencial

Para establecer el paso a paso para evaluar la calidad de los datos y finalmente presentar el dashboard que permita el seguimiento integral al estado de salud de los datos de las organizaciones, conviene estudiar diferentes conceptos relacionados a calidad de datos.

3.1 Marco teórico

En este apartado se describen los componentes teóricos de acuerdo con la investigación, que definen y delimitan conceptos para calidad de datos, de forma general incluyen, referencias internacionales (Normas ISO, Guías), Autores que han sido reconocidos por sus aportes a la calidad de datos.

La calidad de los datos

Según lo descrito por la organización (PowerData, 2022) quienes son un partner del proveedor INFORMATICA LLC uno de los mayores gestores de datos para su análisis, la calidad se define como como las cualidades de un conjunto de datos, que reposan en una base, en algún sistema o en una bodega de datos; y que cumplen con atributos como: “exactitud, completitud, integridad, actualización, coherencia, relevancia, accesibilidad y confiabilidad”; lo anterior con la premisa de que los datos son la base fundamental para tener una buena toma de decisiones dentro de una organización.

Por su parte la compañía IBM en su capítulo de herramientas y soluciones relacionadas con la calidad de los datos menciona: que la calidad de los datos es disposición de la organización, definen los datos de alta calidad, como una herramienta que le permite a las áreas estratégicas de la organización tener una visión integral de toda la compañía.

Adicional a lo ya mencionado es importante indicar que todo el tema de la calidad de los datos viene apalancado por el complimiento de algunas normas internacionales que entregan algunos lineamientos en cuanto a cómo se define si una compañía cumple con los requisitos mínimos de calidad.

La norma (ISO 8000,2016) es una norma Internacional la cual entrega conceptos generales para abordar la gestión de calidad de datos la cual tiene en cuenta la estrategia de calidad, dimensiones de calidad, seguimiento a la calidad de los datos.
Esta norma también habla sobre una gobernanza de datos haciendo las cosas bien, para lo cual tiene en cuenta los siguientes nodos: datos, decisiones y objetivos organizacionales, y puntualmente en el nodo de Datos menciona que estos permiten llegar a una toma de decisiones eficaz y eficiente.
Dicha norma plantea la calidad de los datos como algo sistémico, iniciando con una necesidad de datos, especificación de los datos, analizar el ciclo de vida de los datos, posteriormente la planificación para la mejora, finalmente la ejecución y seguimiento.

Por su parte la norma (ISO/IEC25012, 2021) es una norma internacional que entrega un modelo de calidad el cual utiliza también características de calidad de datos dimensiones que se deben tener en cuenta para evaluar un dato; está compuesta por 15 características, que a su vez se clasifican dentro de 2 grupos:

- Calidad de datos inherente: hace referencia a reglas de negocio que se establecen con los requisitos mínimos de calidad.

- Calidad de datos dependiente del sistema: Corresponde a que la calidad de los datos es recogida y preservada en un sistema informático, es decir la

calidad está correlacionada con el sistema tecnológico en el que los datos se utilizan.

Las Dimensiones que menciona la ISO 25012 que se deben aplicar para la medición de la calidad de los datos son:

- **Exactitud/ Precisión**: se refiere al grado en que los valores de datos representan correctamente la entidad de datos, es decir que datos son incorrectos.
- **Completitud**: indica si todos los datos obligatorios están presentes.
- **Consistencia**: se refiere a la dimensión que permite asegurar que los valores de los datos cumplen con las reglas de contenido y formato.
- **Credibilidad**: Esta dimensión permite definir el grado en que los datos son ciertos y creíbles en el contexto especifico de negocio.
- **Actualidad**: permite revisar el grado en que los datos se encuentran actualizados.
- **Accesibilidad**: establece el grado en que las personas encontrar(acceder) el dato de acuerdo con los permisos específicos.
- **Conformidad**: esta dimensión permite verificar que los datos cumplen con un formato estándar, convención o normativa vigente.
- **Confidencialidad**: está asociada a seguridad de la información y es una dimensión que permite asegurar que los datos solo son accedidos por usuarios autorizados.
- **Eficiencia**: permite analizar el grado en que los datos pueden ser procesados y entregan los niveles de rendimiento esperados.
- **Trazabilidad**: esta dimensión permite analizar si los datos proporcionan un registro de las acciones que los modifican.

- **Comprensibilidad**: los datos pueden ser interpretados y entendidos por cualquier usuario los cuales están expresados en unidades apropiadas y pueden ser leídos.
- **Disponibilidad**: grado que define que los datos pueden ser obtenidos por usuarios o aplicaciones en el momento que lo requieran.
- **Portabilidad**: permite analizar si los datos pueden ser copiados, eliminados o modificados al realizar un cambio de un sistema a otro.
- **Recuperabilidad**: esta dimensión ayuda a comprobar que los datos se mantienen o se pueden recuperar en caso de fallos de un sistema.

Teniendo más claro las definiciones en cuanto a calidad de datos a continuación se aborda alguna literatura referente a documentos que hablan entre otras sobre la gestión y la calidad de los datos.

Fundamentos para el Trabajo con los datos

La Guía de Fundamentos de Gestión de Datos (DMBOOK, 2017) es un documento redactado por más de 120 profesionales en la gestión de datos, dicta unos fundamentos para la gestión de datos además de mencionar los principios y las mejores prácticas en la gestión de los datos.

La guía en su capítulo 1 Gestión de datos, menciona los Datos como un activo vital para la compañía, ya que, entre otras cosas, estos pueden ayudar a dar una visión general de los clientes, productos y servicios, de igual forma pueden impulsar su innovación y a cumplir sus objetivos estratégicos. A pesar de ese entendimiento, actualmente muchas empresas todavía no generan los esfuerzos suficientes para generar valor con sus datos.

Por otra parte, (DAMA-DMBOOK, 2017) en su capítulo 13 Calidad de Datos, dedicado exclusivamente a la calidad de los datos donde mencionan 4 puntos importantes para la gestión de su calidad, los que se describen a continuación:

- Aumentar el valor de los datos de la organización y las oportunidades para usarlos.
- Reducción de riesgos y costos asociados con los datos de baja calidad.
- Aumento de la productividad y la eficiencia de la organización.
- Protección y mejora de la reputación de la organización.

Por su parte (L. Coleman, 2012) en su libro "Data quality measurement for Continuous Improvement" primero menciona que la calidad de los datos se define por el nivel en que cumplen las expectativas de las personas que los consumen, adicionalmente menciona solo 5 dimensiones de calidad de los datos: integridad, puntualidad, validez, consistencia e integridad. Este libro en su apartado de monitoreo y medición de la calidad de los datos sugiere un escenario de evaluación de datos:

- Evaluación inicial de datos.
- Evaluación de proyectos de mejora de calidad de datos.
- Medición continua.

Por otra parte, (D McGilvray, 2008) en el libro "Executing Data Quality Projects" hace un aporte importante al mencionar que no necesariamente los problemas de la calidad de los datos se deben a problemas en los sistemas, si no en una buena parte a errores humanos al momento de su captura, adicionalmente habla de 10 pasos para la estrategia de la calidad de los datos:

1. Definir la necesidad y el enfoque del negocio.
2. Analizar el entorno de la información.
3. Evaluación de la calidad de los datos.
4. Evaluación del impacto comercial.
5. Identificar las causas fundamentales.
6. Definir planes de mejora.
7. Prevenir futuros errores.
8. Sanear los datos actuales.
9. Realizar la implementación de los controles.
10. Comunicar las acciones y los resultados.

De igual forma los autores (D Strong., R Wang., 1996) en su marco de referencia estudiaron los diferentes atributos que pueden intervenir en la calidad de los datos, sin embargo, al ser demasiados, redujeron el numero a 15, las cuales a su vez se sub agruparon en 4 grupos finales.

- Calidad de Datos intrínseca: se busca que el dato cumpla con el objetivo por el cual se creó:
    - Precisión.
    - Objetividad.
    - Credibilidad.
    - Reputación.
- Calidad de datos Contextual: se debe tener en cuenta los objetivos de la organización y para que se van a usar los datos:
    - Valor agregado.
    - Relevancia.
    - Oportunidad.
    - Completitud.

    - Cantidad apropiada de datos.
- Calidad de datos representacional: hace referencia a la fácil lectura e interpretación de los datos:
    - Interoperabilidad.
    - Fácil comprensión.
    - Consistencia representacional.
    - Representación breve.
- Calidad de datos Accesibilidad: corresponde a la facilidad de acceso, sin embargo, también a los niveles de seguridad que se les debe aplicar.
    - Accesibilidad
    - Seguridad para el acceso

Teniendo un poco más entendido la documentación en cuanto a la calidad, es importante mencionar algunos autores que dieron aportes importantes a dicho tema.

Uno de los autores referentes no solo para temas de calidad si no para cualquier proceso en general es **Edwards Deming**, consultor y promotor del concepto de calidad total, famoso por ser el principal impulsor del ciclo DEMMING PHVA (Planear, Hacer. Verificar y Actuar); por lo que es de aclarar que para el alcance de este proyecto solo se van a tener en cuenta las etapas de planear, el hacer y el verificar, en los cuales se identifica el problema y se plantean objetivos para su solución, se desarrolla la propuesta y se verifican los resultados.

Por su parte **(Shewhart, 1931)** con su libro, "Control económico de la calidad de productos manufacturados" estableció los primeros fundamentos para todo el tema del control de la calidad actuales, su principal aporte fue definir los límites de tolerancia que se deben tener para que la calidad sea buena o mala.

Finalmente **(Crosby, 1987) en el libro:** La calidad no cuesta, se centra en que la calidad debe ser una responsabilidad que debe venir desde la alta dirección y propone que se empiece a corregir desde el inicio de cualquier proceso, lo cual a corto plazo le va a reducir costos a la organización.

A continuación, se relacionan algunas consideraciones para la evaluación de la calidad de los datos:

La organización (zipforecasting, 2020) en su artículo: "Evaluación de la calidad de los datos- Métricas y pasos para conocer", para evaluar la calidad de los datos se tiene en cuenta las siguientes métricas: Integridad, validez, oportunidad y consistencia; de igual forma los pasos que sugieren para su evaluación son:

1. Establecer metas a corto y largo plazo.
2. Evaluar fuentes de datos establecidas.
3. Analizar los resultados.
4. Desarrollar métodos de mejora.
5. Implementar soluciones.
6. Supervisar los datos.

Por su parte la empresa (PowerData, 2022) en su artículo:" Calidad de datos. Cómo impulsar tu negocio con los datos", propone 8 dimensiones de calidad: completitud, validez, unicidad, integridad, precisión, coherencia, oportunidad y representación, y además sugiere los siguientes pasos:

1. Descubrimiento de datos.
2. Perfilado de datos.

3. Reglas de calidad.
4. Monitorización de calidad.
5. Reportes de calidad.
6. Corrección de datos.

Adicionalmente a las consideraciones mencionadas, existen algunos análisis que pueden convertirse en alternativas para implementar nuevas formas de medir la calidad.

Por lo anterior, es importante mencionar un método de análisis multicriterio correspondiente al autor (E Cables, 2016) llamado "Método ideal de referencia RIM en la toma de decisiones multicriterio", el cual habla sobre definir el ideal de referencia entre un punto mínimo y un punto máximo, para esto se tienen en cuenta 2 conceptos:

Rango: Es el cualquier intervalo o conjunto de valores.
Ideal de referencia: Corresponde a un conjunto de valores que representan la máxima importancia en un intervalo dado.
De acuerdo con lo descrito por el autor, para aplicar el método RIM se deben seguir los siguientes pasos:

- Paso 1. Definir el contexto de trabajo.
- Paso 2. Obtener la matriz de valoración X, en correspondencia con los criterios definidos.
- Paso 3. Normalizar la matriz de valoración X con el ideal de referencia.
- Paso 4. Calcular la matriz normalizada ponderada.
- Paso 5. Calcular la variación del ideal de referencia normalizado para cada alternativa.
- Paso 6. Calcular el índice relativo de cada alternativa.

- Paso 7. Clasifica las alternativas en orden descendente.

Luego de seguir los pasos anteriores se generará un resultado el cual siempre se encontrará en un rango de 0 a 1, donde entre más cercano el resultado sea a cero más se acercará al ideal de referencia.

Es importante mencionar que para el presente proyecto tenemos una herramienta en la cual ya se encuentran implícitos algunos pasos los cuales se explicarán en el desarrollo de la propuesta.

Luego de revisar algunos aportes a los temas de calidad al igual que algunas consideraciones para su evaluación, a continuación, se mencionan algunas de las herramientas que apoyan la gestión de la calidad de los datos adicionalmente apoyados en la calificación en el cuadrante mágico de Gartner.

Oracle Enterprise Data Quality

Esta herramienta proporciona un entorno de calidad de datos en al cual se abarcan temas como: control de la calidad de los datos, gestión e integración de datos maestros, adicionalmente herramientas de inteligencia de negocios y de migración de datos.

Algunas de sus características clave de EDQ son:

- Auditoria y limpieza de datos.
- Acceso en un ambiente Web.
- Conexiones compatibles con diferentes bases de datos.
- Capacidad de procesamiento de grandes volúmenes de datos.
- Interfaz amigable con el usuario.

- Creación de reglas de validación y transformación de datos,

SQL Server Data Quality Services

(DQS) es una herramienta de calidad de datos la cual dentro de sus principales bondades están: la corrección, enriquecimiento, estandarización y de duplicación de datos, adicionalmente presenta servicios integrados con la nube.

Características Clave:

- Limpieza de datos mediante procesos asistidos.
- Identificació0n de coincidencias basado en reglas.
- Servicios de datos de referencia.
- Elaboración de perfiles para obtener información de los datos en cada una de sus etapas.
- Monitoreo que permite verificar si la solución está dando resultado.
- Base de conocimiento que permite crear procesos de mejora continua.

IBM InfoSphere

El servidor de información de IBM InfoSphere como herramienta de calidad de datos permite realizar limpieza y monitoreo de la calidad de los datos de manera permanente, adicionalmente permite estandarizar, unir y mantener el linaje de datos.

Dentro de sus principales características se resaltan:

- Seguimiento de la Calidad y gestión de datos.

- validación y Estandarización de datos.
- Funciones de clasificación.
- Servicios de Certificaciones.
- Funcionalidad localmente o en la nube.

Talend Data Quality

El programa de calidad de datos de Talend dentro de sus diferentes funcionalidades tiene la posibilidad de creación de perfiles, limpieza y enmascarado de datos en diferentes formatos lo que permite el suministro de datos fiables; Talend Data Quality permite la limpieza de datos mediante técnicas de eliminación de duplicados, validación y normalización por medio de machine learning, adicionalmente ofrece la posibilidad de enriquecimiento de los datos con fuentes externas.

Principales características:

- Rápido desarrollo.
- Fácil mantenimiento.
- Interfaz gráfica.
- Diferentes gamas de conectores y componentes.
- Genera código estándar con java y perl.

A continuación, se mencionan los Sistemas gestores de Bases de datos que se han tenido en cuenta para la ejecución del proyecto.

MySQL

Es un gestor de base de datos relacional de código abierto respaldado por Oracle y que funciona con lenguaje de consulta estructurado (SQL). está disponible para la mayoría de los sistemas operativos más conocidos, como: Linux, UNIX y Windows, dentro de sus beneficios se encuentran:

- Es un gestor de Código abierto.
- Facilidad de uso.
- Compatibilidad con distintos lenguajes de programación y otras bases de datos.
- Se encuentra mucha documentación.
- Seguridad.

SQL Server

Este gestor admite una gran cantidad de aplicaciones y procesamiento de transacciones de inteligencia corporativa y análisis en entornos informáticos empresariales basado en el lenguaje TransactSQL.

Como principales características se encuentran:

- Soporte de transacciones.
- Es una aplicación muy estable.
- Soporte por parte de Microsoft.
- Permite administrar información de otros servidores.

Finalmente se relacionan las herramientas que son más frecuentemente utilizadas para la visualización de datos:

PowerBI

Es una herramienta que permite integrarse con diferentes fuentes de datos, y a su vez contiene una serie de visualizaciones, permite la capacidad de relacionar las

tablas cargadas, adicionalmente tiene la opción de realizar publicaciones en la nube.

Algunas características importantes

- Se deben cargar datos ya procesados ya que maneja un límite de datos.
- Ofrece diferentes servicios de soporte para la versión free y la paga.
- Ofrece más de 3500 opciones para visualización de datos.
- Está diseñada para usuarios principiantes o expertos.
- La interfaz es amigable con el usuario.
- Se concentra más en el modelado de informes y análisis que en el almacenamiento de datos.
- Ofrece la opción de realizar operaciones DAX para formular columnas.
- Actualizaciones mensuales.

Tableau

Es una plataforma de análisis con una interfaz fácil de usar, de igual forma presenta conectividad con datos en la nube.

- Puede manejar una gran cantidad de datos.
- Se pueden utilizar 24 tipos de visualización de datos.
- Presenta un foro comunitario de soluciones.
- Funciona mejor en ambientes de nube.
- Es más utilizado por usuarios avanzados de analítica.
- Es menos intuitivo.
- Solo tiene versión paga.

3.2 Estado del Arte

Teniendo en cuenta que la evaluación es el primer paso en el ciclo de calidad de los datos, esto permite a las organizaciones identificar las diferentes problemáticas que se pueden estar presentando.

Lo anterior con el propósito de planificar acciones que permitan mejorar los datos para garantizar que las diferentes áreas de negocio consumen datos consistentes y son utilizables para la toma de decisiones.

Para abordar este proyecto se ha realizado una investigación de trabajos de grado entorno a la calidad, para de esta forma poder tener un análisis desde diferentes perspectivas, sectores y metodologías de como en Colombia se está implementando la evaluación de calidad, es así como a continuación se presentan diferentes trabajos de autores colombianos.

En un su trabajo (C González, C Hernández, 2017) llamado “diagnóstico de la calidad y el entendimiento de los datos para el análisis y toma de decisiones en las áreas de negocio de la empresa de telecomunicaciones xyzw”. El problema se aborda a raíz de la gran cantidad de reportes que maneja la compañía, de los cuales un 40% presentan reprocesos por temas de la calidad lo que conlleva a unos altos costos en la operación y deficiente gestión de los clientes al no tener la información en los tiempos establecidos.

Por lo anterior a esto plantean un diagnóstico para identificar las principales causas de dichos problemas de calidad, la metodología que proponen es una evaluación de impacto, revisión de alternativas, clasificación del cambio, aprobación del cambio, actualización de planes y documentación, notificación del cambio y la identificación de Stakeholders claves dentro de los procesos.

El método que utilizaron es el Diagnostico de la calidad, recolección de la información, identificación de procesos, seleccionar procesos de análisis, recolección de indicadores, recolección de datos, realización de Focus Group en el cual se reúnen a diferentes usuarios y se reciben sus percepciones en cuanto al proceso, visitas de campo, entrevistas a usuarios, análisis de resultados, presentación de propuestas, seguimiento a la implementación, revisión plan de mejora.

Finalmente se concluye la propuesta en un enfoque en la mejora de algunos procesos orientados por la gestión de la calidad, sin embargo, no tienen en cuenta las dimensiones de la calidad de los datos.

Por otra parte, (D Rodríguez, 2020), presenta el trabajo “desarrollo RPA para monitoreo de calidad de datos y generación de alertas”, en el cual plantean como la utilización de la metodología tdqm (Total Data Quality Management) cuyo problema consiste en que no se tiene un método que genere alertas al momento en que se presente un problema de calidad en algún cargue de información. En este sentido se centran en definir los datos, medir y analizar su calidad, y optimizar el proceso de calidad de los datos apoyados en las dimensiones: Precisión, completitud, y consistencia.

Por lo que como estrategia proponen definir unas funciones y responsabilidades sobre la gestión de los datos, adicionalmente realizan algunas comparaciones en cuanto a algunas herramientas de calidad, presentan algunas gráficas con mediciones de la calidad de los datos.

En la propuesta del proyecto mencionan pilares de la calidad de los datos como: gobierno de datos, seguridad, integridad, almacenamiento, arquitectura, bases de datos y datos maestros; adicionalmente crean unas reglas de negocio para generar

alertas en el momento en que el cargue de algún archivo presentaba algún error en la calidad, sin embargo, no da un nivel de importancia a cada una de las dimensiones o de los atributos trabajados.

Por su parte (M Rodríguez,2019), menciona en su trabajo de “Plan de gestión de calidad de datos para mejorar la oportunidad y pertinencia de la información de la oferta institucional en la dirección de apropiación del ministerio TIC” de la universidad Externado. Centran el problema en tener un monitoreo sobre la calidad, homologación y estandarización de los datos, la propuesta se basa en, realizar un diagnóstico a través de un muestreo de datos, y posteriormente la aplicación de una entrevista a los responsables del manejo de los datos. Finalmente presentan un muestreo de los hallazgos en cuanto a la calidad de los datos.

Este trabajo se enfoca en la medición del volumen de datos, cantidad de devoluciones, diversidad de errores e importancia misional. Finalmente genera una matriz en la cual se muestran los fallos detectados en calidad en alguno de los atributos del set de datos. Por cada uno de los atributos dice si la calidad de ese registro es Baja, media o alta, sin embargo, no menciona a nivel de detalle cuales fueron las herramientas utilizadas para dicha medición.

Impacto

Al realizar una evaluación integral de la calidad de los datos, se proporcionará un mayor conocimiento a la organización en cuanto al estado global que tienen sus datos a partir del análisis de los problemas existentes en los atributos críticos y las dimensiones aplicadas de acuerdo con sus necesidades de negocio, de igual forma por medio del dashboard se entrega una herramienta que automatiza la consolidación de resultados orientada a tomar acciones preventivas y correctivas para mejorar los datos y de esta manera generar valor para la toma de decisiones.

Componente de Innovación

El componente de innovación está en que se realiza una evaluación integral de la calidad no solo a nivel de dimensiones individuales si no de todas las utilizadas para llegar un solo resultado, esto apoyado en la aplicación del método RIM, que se convierte en una herramienta muy útil para apoyar el proceso de toma de decisiones evaluando diferentes alternativas dadas, lo cual entrega a la organización el estado actual de la calidad de los datos de una forma integral, valorando las dimensiones en el orden importancia que le genera más impacto al negocio y definiendo un umbral específico para el dato.

4.**Metodología**

Partiendo del contexto anterior, la metodología aplicada para el desarrollo de este proyecto está basada en las buenas prácticas para la gestión de calidad de datos que entrega la Guía de Fundamentos de Gestión de datos DMBOK del DAMA en el capítulo de calidad de datos, donde es importante resaltar que calidad de datos es un área de conocimiento fundacional dentro de la gestión de datos.

Si bien, el propósito de calidad de datos desde la perspectiva de DMBOK2 y la norma internacional ISO 8000 para calidad de datos aborda actividades de planificación, implementación y control donde se usan técnicas de gestión de calidad de datos para garantizar que los datos que se entregan a la organización son idóneos para su uso. El enfoque del presente proyecto está delimitado en la etapa evaluación de la calidad de los datos, la cual permite entender el estado de los datos de la organización para cuantificar la calidad de los datos y determinar los esfuerzos que se requieren para la prevención y mejora de estos.

Una vez se ha definido el alcance de la metodología, se establecen los pasos que se deben tener en cuenta para la evaluación de calidad de los datos en una organización y las buenas prácticas para el desarrollo de esta con base en lo definido por la Guía de Fundamentos de Gestión de Datos DMBOK del DAMA, plasmando el propósito, las actividades, técnicas que pueden ser utilizadas para obtener mejores resultados.

Los pasos clave para evaluar la calidad de los datos se ilustran a continuación:

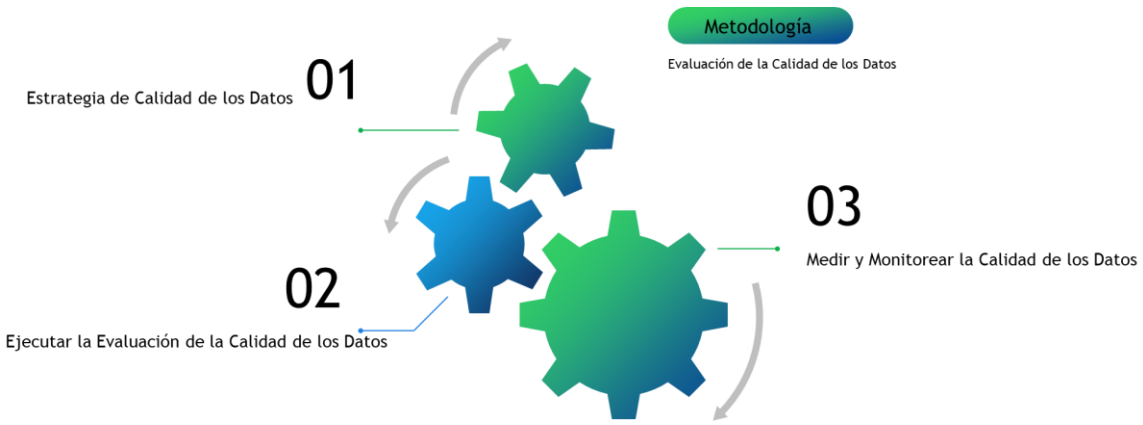


Ilustración 1 Pasos clave para Evaluar la Calidad de los Datos. Fuente: Elaboración Propia adaptado del DAMA DMBOK2

De acuerdo con la Figura anterior, se describen a continuación cada uno de los pasos para evaluar la calidad de los datos en una organización, donde se deben ejecutar en la secuencia presentada:

a) **Estrategia de Calidad de los Datos.** Precisar la Estrategia de Calidad de Datos es el primer paso que se debe considerar para Evaluar la Calidad de los datos en una organización, básicamente consiste en realizar un entendimiento de las necesidades del negocio a través de un caso de uso en torno a la calidad de los datos, se definen los objetivos y se delimita el alcance de la evaluación de calidad de datos.

   Teniendo definido el caso de uso, el objetivo y el alcance de la evaluación, se procede a identificar elementos de datos críticos. Los elementos de datos críticos son reconocidos y seleccionados por cada organización, ya que los esfuerzos para la evaluación de calidad de datos deben centrarse en los datos más importantes para el negocio ya sea porque impactan a nivel financiero, operativo, regulatorio o reputacional.

Luego de contar con los elementos de datos críticos seleccionados, se procede a reconocer y determinar las fuentes de datos donde residen estos elementos, no sólo basta con determinar la fuente de datos, también se hace necesario especificar las tablas en la base de datos, ya que con esta información se pude documentar los metadatos técnicos cómo: fuente de datos, nombre de la columna en la tabla, llaves, tipo de dato y longitud. Es relevante mencionar que la organización debe establecer una muestra representativa del conjunto de datos a evaluar y determinar los mecanismos de acceso a los datos según la necesidad, es decir, si autoriza acceso por conexión de red privada virtual o VPN (Virtual Private Network) y otorgar privilegios de acceso a los datos de acuerdo con la sensibilidad de los datos almacenados en la fuente de datos.

Para realizar la evaluación de la calidad de los datos se debe considerar el componente tecnológico, ya que las herramientas de calidad de datos proporcionan funcionalidades para el perfilamiento de datos, entregan un conjunto de reglas predefinidas aplicables a los elementos de datos y permiten crear reglas propias de acuerdo con las necesidades del negocio.

b) **Ejecutar la Evaluación de la Calidad de los Datos.** Partiendo del caso de uso el cual describe las necesidades más complejas que el negocio desea resolver se inicia el perfilamiento de los datos a las fuentes de datos especificadas, este perfilamiento se realiza con la herramienta de calidad de datos seleccionada, y corresponde a un tipo especifico de análisis de datos, el cual permite descubrir características en los datos cómo patrones existentes en los datos, estructuras de los datos e identificar anticipadamente nulos existentes y valores atípicos.

A partir del análisis preliminar se establecen las dimensiones de calidad de datos aplicables a los elementos de datos críticos seleccionados; las dimensiones de calidad de datos le permiten al rol responsable de calidad de datos caracterizar las reglas de calidad para evaluar los datos, medir la idoneidad de los datos para su uso. la palabra dimensión significa en calidad de datos aquellos aspectos de los datos que se pueden medir y a través de los cuales se puede cuantificar la calidad de los datos de una organización. Las dimensiones de calidad de datos seleccionadas para el presente ejercicio son:

- **Completitud**. La completitud permite identificar si están todos los registros presentes a nivel de conjunto de datos, columnas o atributos y que además son necesarios para el negocio. En esta dimensión es clave identificar si el elemento de dato es obligatorio, opcional o inaplicable.
- **Conformidad**. La conformidad valida si los datos que están en las columnas o atributos de una tabla están en el formato estándar.
- **Consistencia**. Los datos son coherentes y no presentan contradicciones.
- **Duplicidad.** Permite identificar registros duplicados.
- **Validez.** La validez es una dimensión que permite comparar los datos contra un rango de valores como una tabla de referencia.

Una vez se comprenden las características en los datos con el perfilamiento de datos y se determina las dimensiones de calidad de datos, es importante profundizar en los detalles de cada uno de los elementos de datos críticos, es decir realizar el levantamiento de reglas de negocio aplicables a cada uno, este paso se realiza con entrevistas a los involucrados relevantes para documentar como deben existir los datos en las fuentes de datos para que sean utilizables dentro de la organización. A continuación, algunas de las preguntas que se

pueden realizar para el levantamiento de reglas para evaluar la calidad de los datos:

- ¿Qué elementos de datos son obligatorios y cuales son opcionales?
- Cuál es el valor asignado a un elemento si debe estar dentro de un rango determinado.
- Se revisa si los valores de los datos se deben comparar contra un valor en una tabla de referencia.
- Se determina qué entidades de datos deben tener una representación única.
- En que formato deben venir los datos.

Seguidamente de documentar las reglas para evaluar la calidad de los datos, se pasa al diseño e implementación de dichas reglas, este proceso depende de la herramienta seleccionada para el ejercicio, las reglas para evaluar la calidad de los datos se pueden diseñar e implementa con el apoyo de un lenguaje SQL o expresiones regulares. Tras la implementación de las reglas para evaluar la calidad de los datos, se visualizan los niveles de no conformidad de los datos, es decir, se documentan los hallazgos de una forma clara para facilitar el entendimiento del estado de los datos.

Por lo general, después de identificar los problemas en los datos, se realiza el análisis de causa raíz; el análisis de causa raíz es un método de resolución de problemas que permite descubrir los factores que contribuyen en la generación de los problemas en los datos, con el fin de identificar soluciones apropiadas; es clave mencionar que para ejecutar esta actividad se requiere la participación de las personas de negocio y TI. El rol designado para calidad de datos lidera y facilita la actividad, pero el éxito radica en la colaboración

interfuncional. Existen diferentes métodos para realizar el análisis de causa raíz, sin embargo, para el desarrollo del presente proyecto se aplicará la herramienta espina de pescado o también conocida como causa – efecto, donde se abordan causas raíz técnicas y no técnicas.

c) **Medir y Monitorear la Calidad de los Datos.** Las organizaciones necesitan contar con la capacidad de medir y monitorear los datos, por tal razón, debe definir los umbrales de calidad de datos, los cuales son la presentación numérica del límite aceptable de medida para los datos con problemáticas. Cuando una de las medidas es mayor que el umbral establecido indica un problema de calidad de datos o un cambio significativo que requiere ser investigado. Los umbrales de calidad de datos se pueden establecer basado en niveles razonables que determine el negocio.

   Producto de la investigación para el desarrollo del presente ejercicio y con el propósito de entregar a las organizaciones una evaluación integral de la calidad de los datos, se ha adoptado el método del ideal de referencia (RIM) desarrollado por Cables et al. (Cables, et al, 2016), ya que este método forma parte de los métodos de análisis multicriterio que apalancan la toma de decisiones. Es relevante destacar que el método (RIM) inicia con la identificación del ideal de referencia para cada uno de los criterios que serán utilizados para evaluar cada alternativa. Se basa principalmente en dos conceptos:

   - **Rango**. Es cualquier intervalo, conjunto de etiquetas o un conjunto simple de valores pertenecientes a cualquier universo de discurso.

- **Ideal de referencia.** Es un intervalo, conjunto de etiquetas o valores simples que representan la máxima importancia o relevancia en un Rango dado.

Para utilizar el método RIM, se debe seguir los siguientes pasos (Cables, et al, 2016):

- **Descripción de los criterios.** Aquí el método (RIM) utiliza un vector de pesos para indicar la importancia relativa entres las dimensiones de calidad de datos seleccionadas. Para la determinación del vector de pesos, en este proyecto se utilizará el método propuesto por Cables y Lamata (Lamata y Cables, 2009) por ser de fácil implementación y utilizar la función de comportamiento lineal donde la distancia entre pesos adyacentes es constante. La función para el cálculo del vector de pesos a utilizar es la propuesta por Borda-Kendall:

$$w_i = \frac{2(n+1-i)}{n(n+1)}$$

Donde:

- n, es la cantidad de elementos del vector de pesos.
- $i = 1, 2, \ldots, n$, donde cada i corresponde a un peso o ponderación.

Después de aplicar el método (RIM) se entregará la puntuación integral del estado de salud de los datos de la organización por elementos de datos críticos. Para facilitar el análisis de datos se debe construir un dashboard en Power BI que consolida los resultados de la evaluación de calidad de los datos.

5. Desarrollo de la propuesta

En este capítulo se presenta el desarrollo de la propuesta para elaborar un Dashboard que permitirá a las organizaciones de múltiples sectores realizar seguimiento integral al estado de salud de los datos en las organizaciones.

a) Estrategia de Calidad de los Datos

En esta actividad se simuló un caso de uso de una organización especializada en seguros de vida y riesgos laborales. Para la construcción del caso de uso en torno a la calidad de los datos se detalla la problemática de datos junto con los beneficios que obtendrá a organización al establecer y realizar seguimiento a la evaluación de la calidad de los datos. la alineación con los objetivos estratégicos y las áreas de negocio impactadas.

A continuación, se describe brevemente el problema de datos:

- En la organización de seguros de vida y riesgos laborales, el área comercial de seguros de personas, donde se encuentra la gerencia de mercadeo masivo realizan campañas para ofrecer a los clientes nuevos productos, pero actualmente tienen problemas para contactar a los clientes por datos faltantes e incompletos, lo que genera sobrecosto al generar las campañas de email marketing y reprocesos por la no entrega oportuna de comunicaciones con información relevante de cambios y actualizaciones en condiciones de sus pólizas. La baja calidad de los datos impacta igualmente a otras áreas del negocio como el área de operaciones ya que les genera reprocesos porque los clientes no reciben oportunamente las comunicaciones respecto a las actualizaciones o cambios en las condiciones de las pólizas. Al realizar la evaluación de calidad de los datos, podrá identificar y analizar las problemáticas para lograr procesos de negocio más

efectivos, y aumentará la confianza en los datos y así como mejorar el relacionamiento con el cliente ya que aumentará el contacto efectivo con los clientes reduciendo devoluciones de las comunicaciones remitidas de forma física, electrónica y telefónicamente.

Teniendo en cuenta el caso presentado anteriormente, se establece el objetivo y alcance de la evaluación de la calidad de datos, en esta actividad se identifican la del conjunto de datos de CLIENTES y los elementos de datos críticos objeto de la evaluación, la fuente de datos donde residen estos datos y la herramienta de calidad de datos a utilizar; una fuente de datos puede ser una base de datos o una hoja de cálculo. La fuente de datos inicial utilizada para el presente ejercicio es una hoja de cálculo en Excel creada con datos dummies de CLIENTES el cual contiene 100.000 registros y 31 elementos de datos de clientes. Para dar un mayor entendimiento a continuación se presenta el diccionario de datos del set de datos de CLIENTES:

Atributo	**Nombre_Campo_Negocio**	**Descripción**
1	id_cliente	Id dentro de la base de datos que identifica al cliente como un único registro
2	Tipo_de_documento	Tipo de Documento de Identificación del cliente
3	Numero_de_Documento	Numero de Documento de Identificación del cliente
4	Fecha_de_Nacimiento	Fecha de Nacimiento del cliente
5	Fecha_Expedicion_Documento	Fecha de Expedición Documento
6	Genero	Genero
7	primer_ape	Primer Apellido
8	segundo_ape	Segundo Apellido
9	primer_nombre	Primer Nombre
10	segundo_nombre	Segundo Nombre

Atributo	Nombre_Campo_Negocio	Descripción
11	Estado_Civil	Estado Civil
12	Direccion	Dirección de residencia del cliente
13	Cod_Ciudad	Código Ciudad de acuerdo con la clasificación del Dane
14	Ciudad	Ciudad de acuerdo con la clasificación del Dane
15	Cod_Depto	Código Departamento de acuerdo con la clasificación del Dane
16	Departamento	Departamento de acuerdo con la clasificación del Dane
17	Codigo_dane	Código Dane de acuerdo con la clasificación del Dane
18	Telefono_1	Teléfono (Fijo)
19	Telefono_2	Teléfono (Celular)
20	Email	Email
21	Fecha_Ingreso_Compania	Fecha ingreso a la compañía
22	Tipo_de_Persona_N_J	Tipo de Persona (Natural o Jurídica)
23	Profesion	Profesión
24	Ocupacion	Ocupación
25	CIIUEmpresas	CIIU(Empresas) corresponde a una forma más detallada de clasificar las actividades Económicas.
26	Descripcion_CIIUEmpresas	Descripción CIIU(Empresas)
27	Ingresos_Mensuales	Ingresos Mensuales
28	pagina_de_internet	Página de Internet, la página Web de la Compañía, solo aplica para personas jurídicas.
29	Estado_Actual_en_la_Compania	Estado Actual en la Compañía, si está Activo o Anulado
30	Cantidad_Productos	Nivel de Riesgo, clasificación interna de acuerdo con criterios como edad, ubicación geográfica, entre otras.

Atributo	Nombre_Campo_Negocio	Descripción
31	Ramo	Ramo al que pertenece el producto del cliente

Tabla 2: diccionario de datos. Fuente Propia

La distribución de los tipos de campo se encuentra de la siguiente forma:

Tipo de Campo	Cantidad
Bigint	2
DateTime	3
Int	7
Varchar	19
Total general	31

Tabla 3: tipos de datos set. Fuente propia

Para identificar los elementos de datos más importantes del conjunto de datos de CLIENTES, se representó que puede ser más importante para el proceso de negocio teniendo en cuenta el impacto que genera la baja calidad de los datos desde (4) cuatro perspectivas: Financiero, Operativo, Regulatorio y Reputacional.

A continuación, se presenta el impacto de la baja calidad de datos para el negocio:
Área de Negocio. Gerencia de Mercadeo masivo
Proceso de Negocio. Comercialización de productos y servicios
Elementos de datos críticos: 13

Elemento de Dato Crítico (CDES)	Impacto para el negocio			
	Financiero	**Operativo**	**Regulatorio**	**Reputacional**
Tipo de documento	N/A	N/A	N/A	N/A
Número Documento	Es importante para la remisión de información exógena a los entes de control	Para poder tener información completa en cuanto a asegurados y beneficiarios	Porque se debe tener la información completa para reportar a entes de control en los cuales se evidencia la información completa y confiable	La información debe ser integra y debe tener niveles de confidencialidad y de tratamiento en los cuales no se vea afectados o expuestos los datos del cliente
Fecha Nacimiento	N/A	Para cálculos de tasas y primas	Para soportar los cálculos actuariales que se realizan sobre los asegurados	Para tener información que permita identificar la identidad de los clientes

Elemento de Dato Crítico (CDES)	Impacto para el negocio			
	Financiero	**Operativo**	**Regulatorio**	**Reputacional**
Fecha Expedición	N/A	Para cálculos de tasas y primas	Para soportar los cálculos actuariales que se realizan sobre los asegurados	Para tener información que permita identificar la identidad de los clientes
Primer Apellido	Para reporte de información exógena a entes de control	N/A	Para generar comunicaciones y tener control sobre las solicitudes	Para tener información que permita identificar la identidad de los clientes
Primer Nombre	Para reporte de información exógena a entes de control	N/A	Para generar comunicaciones y tener control sobre las solicitudes	Para tener información que permita identificar la identidad de los clientes
Dirección	N/A	Para generar comunicaciones y tener control sobre las solicitudes	Para generar comunicaciones y tener control sobre las solicitudes	Para generar comunicaciones y tener control sobre las solicitudes
Código Ciudad	N/A	Para generar comunicaciones y tener control sobre las solicitudes	Para generar comunicaciones y tener control sobre las solicitudes	Para generar comunicaciones y tener control sobre las solicitudes
Código Departamento	N/A	Para generar comunicaciones y tener control sobre las solicitudes	Para generar comunicaciones y tener control sobre las solicitudes	Para generar comunicaciones y tener control sobre las solicitudes
Email	N/A	Para poder enviar y responder comunicaciones	Para poder enviar y responder comunicaciones	Para poder enviar y responder comunicaciones

Elemento de Dato Crítico (CDES)	Impacto para el negocio			
	Financiero	**Operativo**	**Regulatorio**	**Reputacional**
Telefono1	N/A	Para generar insumos para estrategias comerciales	Para tener datos actualizados d ellos clientes	Para tener contacto en caso de generar alguna información importante.
Telefono2	N/A	Para generar insumos para estrategias comerciales	Para tener datos actualizados d ellos clientes	Para tener contacto en caso de generar alguna información importante.

Tabla 4 Impacto para el negocio. Fuente: Adaptado del material de DAMA Colombia

Cabe mencionar que cada organización puede identificar los elementos de datos críticos y establecer el impacto de la baja calidad de los datos para su negocio.

b) Ejecutar la Evaluación de la Calidad de los Datos.

Una vez establecido el caso de uso, los elementos de datos críticos. la fuente de datos y el conjunto de datos objeto de evaluación, se inicia la ejecución de la evaluación de calidad de datos.

Por ser un ejercicio de carácter académico y entregar un mayor entendimiento de la evaluación, se presenta el flujo del proceso para ejecutar la evaluación de calidad de los datos que inicia con el cargue del set de datos hasta la generación del Dashboart es el siguiente:

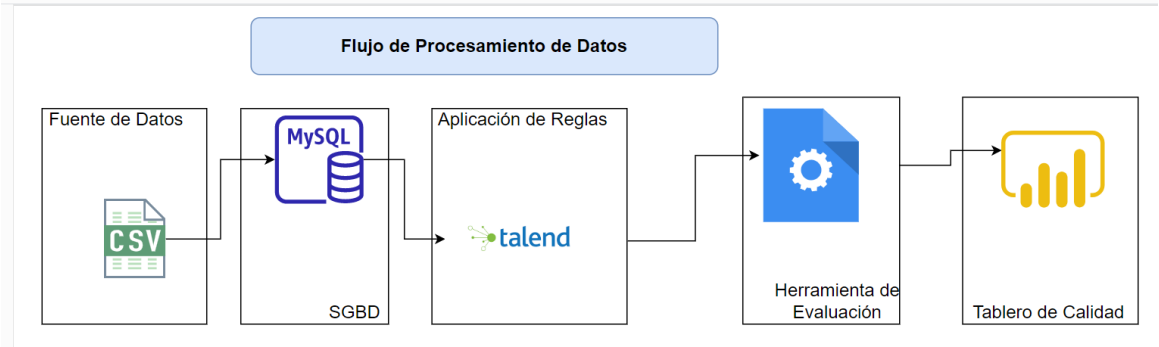


Ilustración 2: Flujo procesamiento de datos. Fuente propia

A continuación, se puede visualizar la creación de la base de datos MySQL y el cargue del Set de Datos de CLIENTES:

Cargue en MySQl

Primero se realiza la creación de la tabla con los atributos mencionados:

```
create table Clientes (
id_cliente Int(6) primary key,
Tipo_de_documento Varchar(4),
Numero_de_Documento Int(10),
Fecha_de_Nacimiento Datetime,
Fecha_Expedicion_Documento Datetime,
Genero Varchar(1),
primer_ape Varchar(50),
segundo_ape Varchar(50),
primer_nombre Varchar(650),
segundo_nombre Varchar(50),
Estado_Civil Varchar(12),
Direccion Varchar(80),
Cod_Ciudad Int(3),
Ciudad Varchar(30),
Cod_Depto Int(2),
Departamento Varchar(20),
Codigo_dane Int(5),
Telefono_1 Int(7),
Telefono_2 BIGINT(10),
```

```
Email Varchar(50),
Fecha_Ingreso_Compania Datetime,
Tipo_de_Persona_N_J Varchar(1),
Profesion Varchar(120),
Ocupacion Varchar(120),
CIIUEmpresas varchar(6),
Descripcion_CIIUEmpresas Varchar(200),
Ingresos_Mensuales BIGINT(20),
pagina_de_internet Varchar(30),
Estado_Actual_en_la_Compania Varchar(10),
Cantidad_Productos int(2),
Ramo varchar (25));
```

Posteriormente por medio del procedimiento LOAD DATA INFILE se realiza el cargue del set completo de datos:

```
LOAD DATA INFILE 'C:\\ProgramData\\MySQL\\MySQL Server
8.0\\Uploads\\SetDatos.txt'
INTO TABLE cargue2
CHARACTER SET latin1
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\r\n'
(id_cliente,Tipo_de_documento,Numero_de_Documento,Fecha_de_Nacimiento,F
echa_Expedicion_Documento,Genero,primer_ape,segundo_ape,primer_nombre,s
egundo_nombre,Estado_Civil,Direccion,Cod_Ciudad,Ciudad,Cod_Depto,Departam
ento,Codigo_dane,Telefono_1,Telefono_2,Email,Fecha_Ingreso_Compania,Tipo_d
e_Persona_N_J,Profesion,Ocupacion,CIIUEmpresas,Descripcion_CIIUEmpresas,Ingr
esos_Mensuales,pagina_de_internet,Estado_Actual_en_la_Compania,Cantidad_Pr
oductos);
```

La siguiente ilustración permite evidenciar el cargue de los registros a la base de datos:

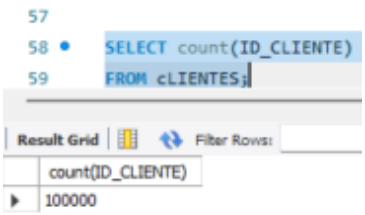
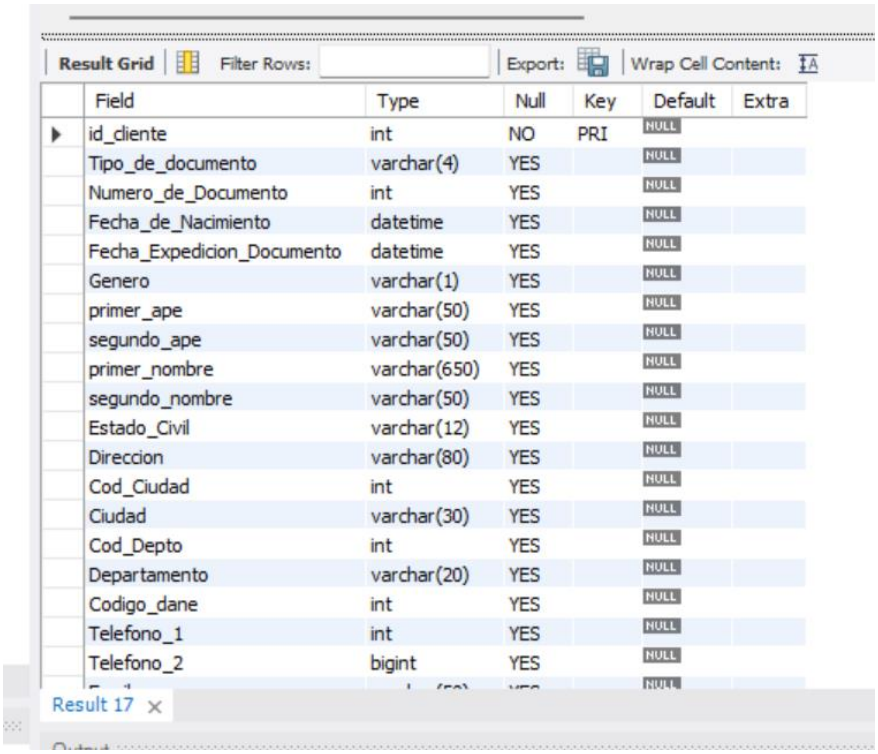


Ilustración 3: Conteo registros cargados. Fuente propia

El set de datos cargado se visualiza en la siguiente ilustración:



Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Field	Type	Null	Key	Default	Extra
id_cliente	int	NO	PRI	NULL	
Tipo_de_documento	varchar(4)	YES		NULL	
Numero_de_Documento	int	YES		NULL	
Fecha_de_Nacimiento	datetime	YES		NULL	
Fecha_Expedicion_Documento	datetime	YES		NULL	
Genero	varchar(1)	YES		NULL	
primer_ape	varchar(50)	YES		NULL	
segundo_ape	varchar(50)	YES		NULL	
primer_nombre	varchar(650)	YES		NULL	
segundo_nombre	varchar(50)	YES		NULL	
Estado_Civil	varchar(12)	YES		NULL	
Direccion	varchar(80)	YES		NULL	
Cod_Ciudad	int	YES		NULL	
Ciudad	varchar(30)	YES		NULL	
Cod_Depto	int	YES		NULL	
Departamento	varchar(20)	YES		NULL	
Codigo_dane	int	YES		NULL	
Telefono_1	int	YES		NULL	
Telefono_2	bigint	YES		NULL	

Result 17

Ilustración 4: set de datos cargado. Cuente propia

Se simula la selección de dimensiones aplicables a los datos y se determina cuales se aplican a cada elemento de datos critico:

Elemento de Dato critico	**Dimensiones de Calidad de Datos**				
	Completitud	**Conformidad**	**Consistencia**	**Duplicidad**	**Validez**
Ciudad	X		X	X	
Departamento	X		X	X	
Dirección	X			X	
Email	X	X	X	X	
Fecha Expedición	X			X	X
Fecha Nacimiento			X	X	X

Numero Documento	X			X	
Primer Apellido	X	X	X	X	
Primer Nombre	X	X	X	X	
Telefono1	X	X	X	X	X
Telefono2	X	X	X	X	X
Tipo Documento	X		X		

Tabla 5: Elementos de Datos y Dimensiones. Fuente propia

Las dimensiones seleccionadas para el ejercicio son:

- **Completitud**. La completitud permite identificar si están todos los registros presentes a nivel de conjunto de datos, columnas o atributos y que además son necesarios para el negocio. En esta dimensión es clave identificar si el elemento de dato es obligatorio, opcional o inaplicable.
- **Conformidad**. La conformidad valida si los datos que están en las columnas o atributos de una tabla están en el formato estándar.
- **Consistencia**. Los datos son coherentes y no presentan contradicciones.
- **Duplicidad.** Permite identificar registros duplicados.
- **Validez.** La validez es una dimensión que permite comparar los datos contra un rango de valores como una tabla de referencia.

El acceso a los datos se realiza directamente desde Talend Open Studio a una Base de Datos MySQL de acuerdo con los permisos de seguridad de acceso establecidos por la organización:

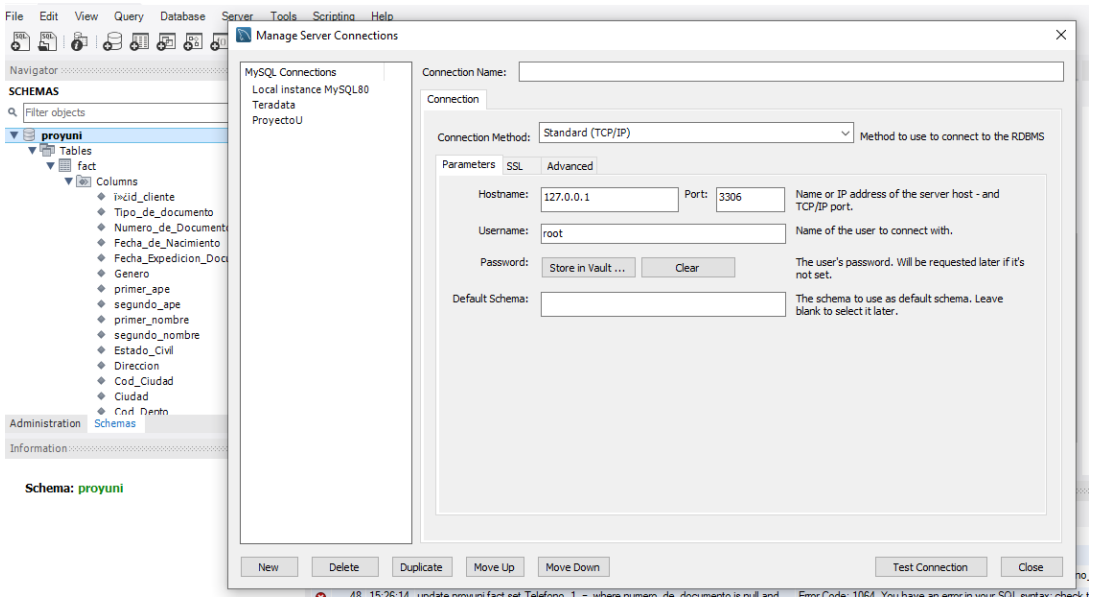


Ilustración 5: Acceso a los datos

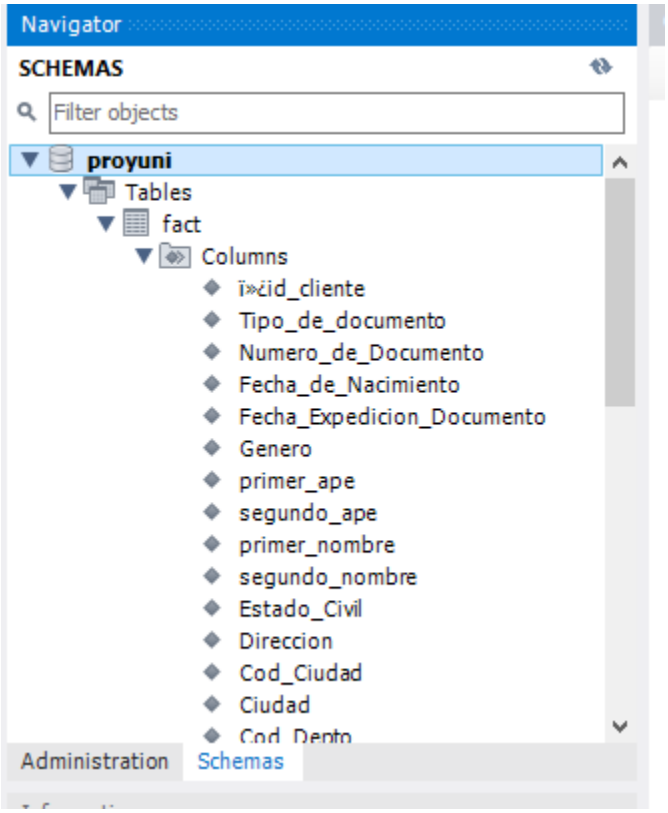


Ilustración 6: Configuración del set de datos en Talend Open Studio

La siguiente ilustración permite visualizar el set de datos cargado en MySQL:

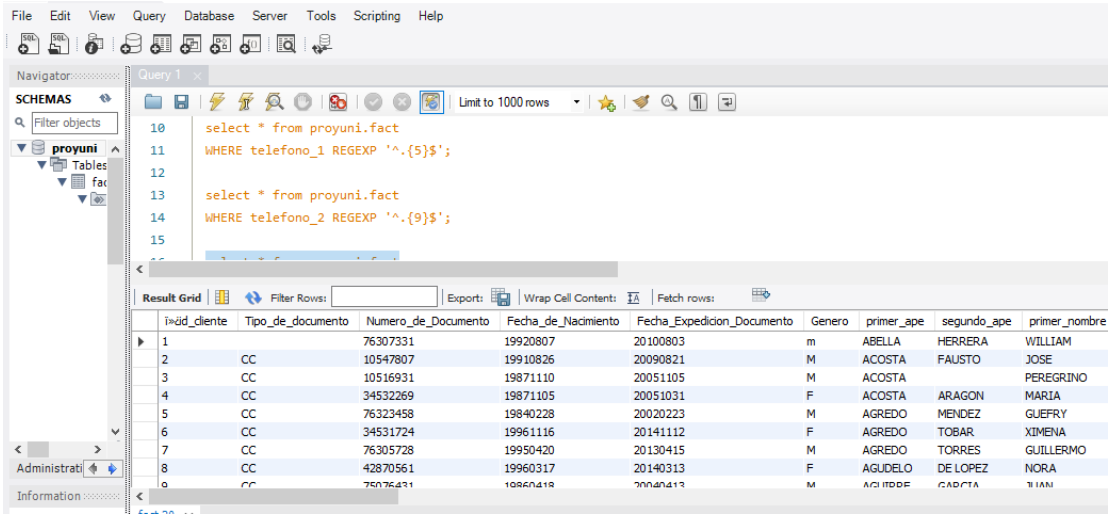


Ilustración 7: Set de datos cargado en MySQL

Al contar con los datos cargados en la base de datos MySQL, se procede a crear el proyecto en la herramienta para realizar el perfilamiento de los datos:

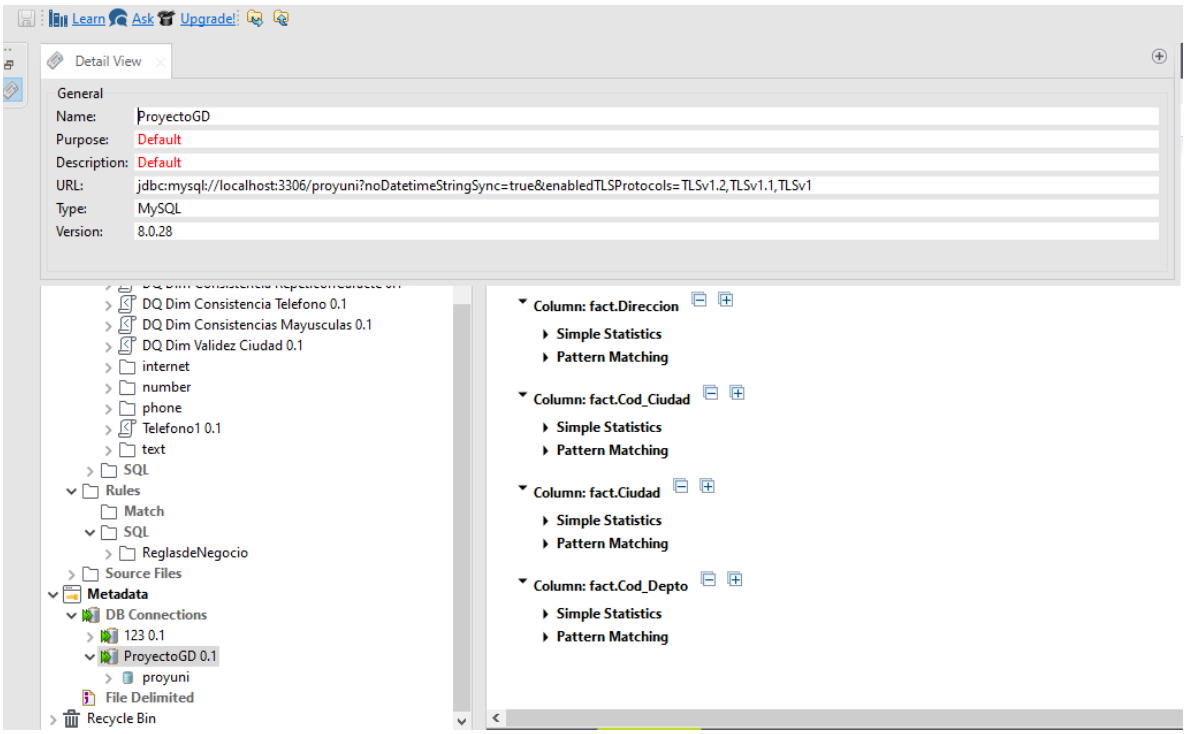


Ilustración 8: Creación del proyecto para evaluar la calidad de datos

Se inicia con la configuración de las reglas para evaluar la calidad de los datos en la herramienta a través del uso de expresiones regulares:

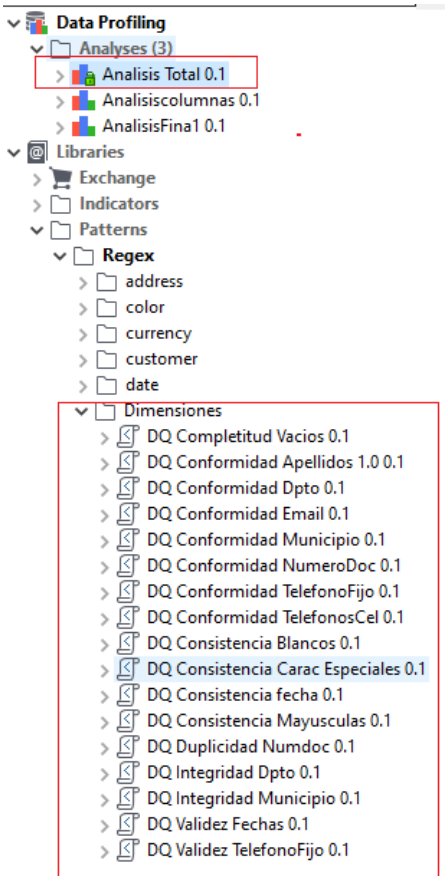


Ilustración 9: Configuración de reglas para evaluar la calidad de los datos

Haciendo uso de las expresiones regulares "REGEXP_LIKE(NOMBRES, '[Ññ!"#$%&/()?¡¿*:;|{}^[<>`Ã]')" se procede a implementarlas en la herramienta de calidad de datos:

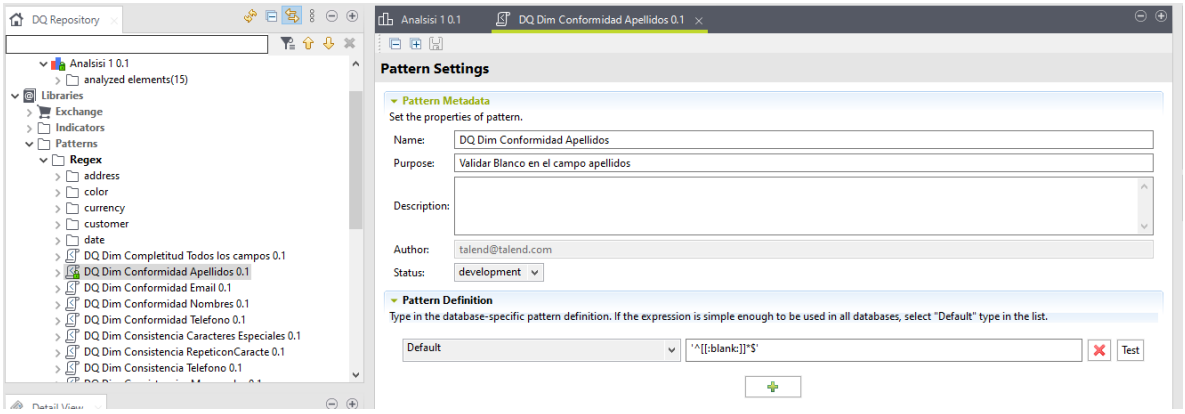


Ilustración 10: Configuración de reglas para evaluar la calidad de los datos a través de expresiones regulares

Las siguientes ilustraciones permiten visualizar una muestra de las reglas para evaluar la calidad de los datos implementadas:

Verifica caracteres especiales.

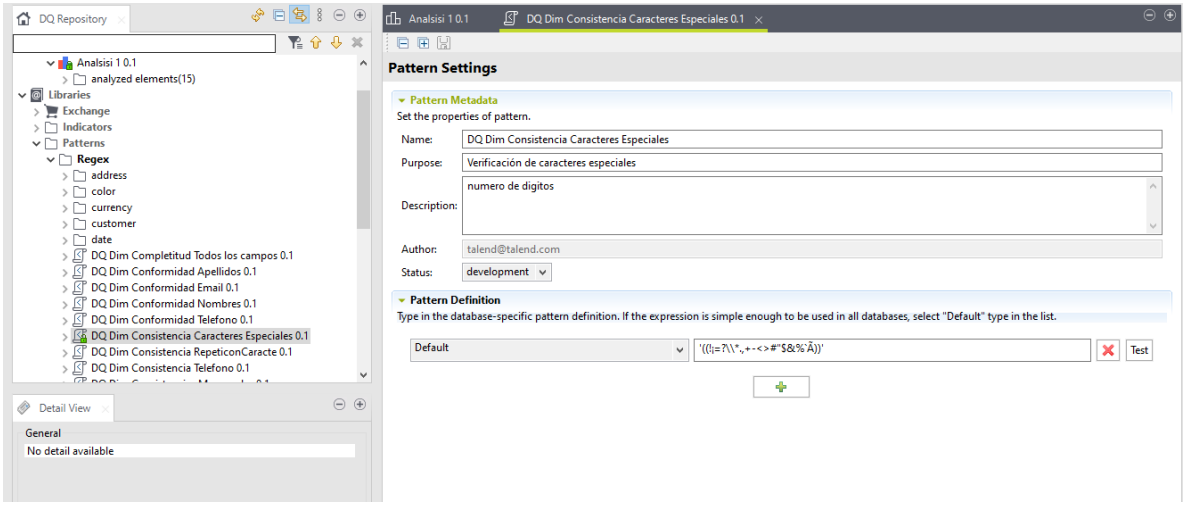


Ilustración 11: Configuración de reglas para evaluar la calidad de los datos a través de expresiones regulares

Válida que no exista textos en mayúscula.

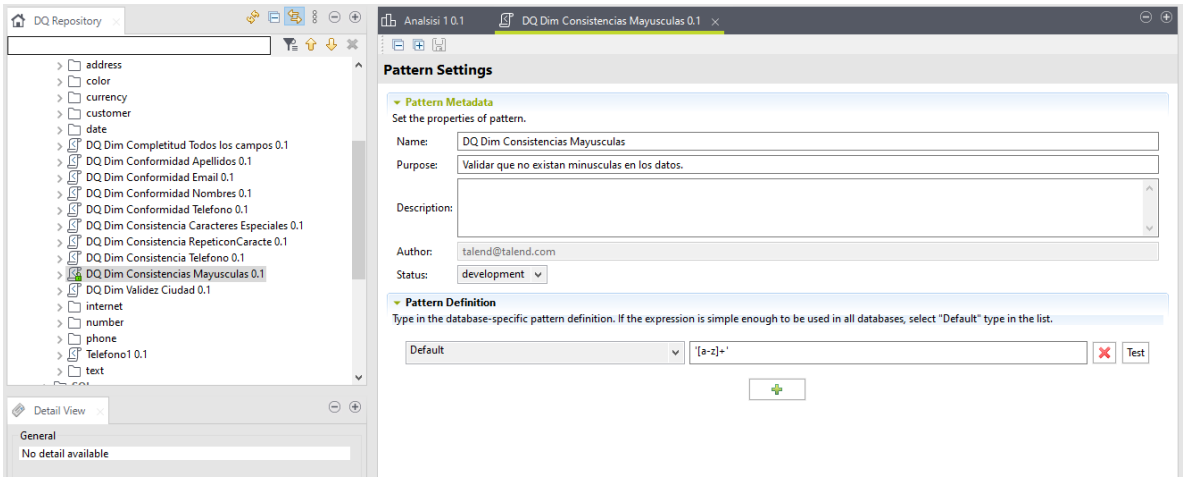


Ilustración 12: Válida que no exista textos en mayúscula

Valida la Conformidad Email.

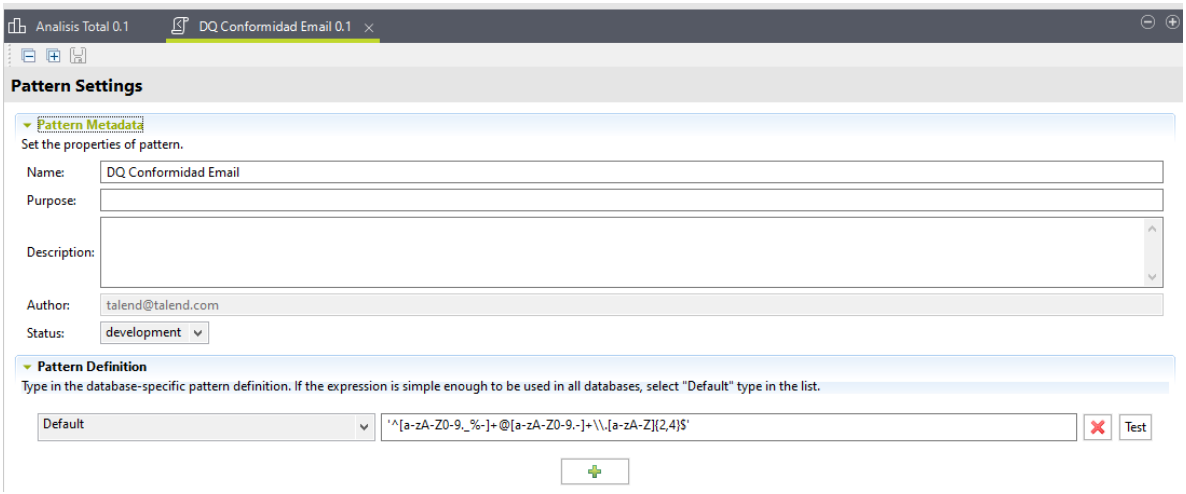


Ilustración 13: Valida la Conformidad Email

Conformidad municipio.

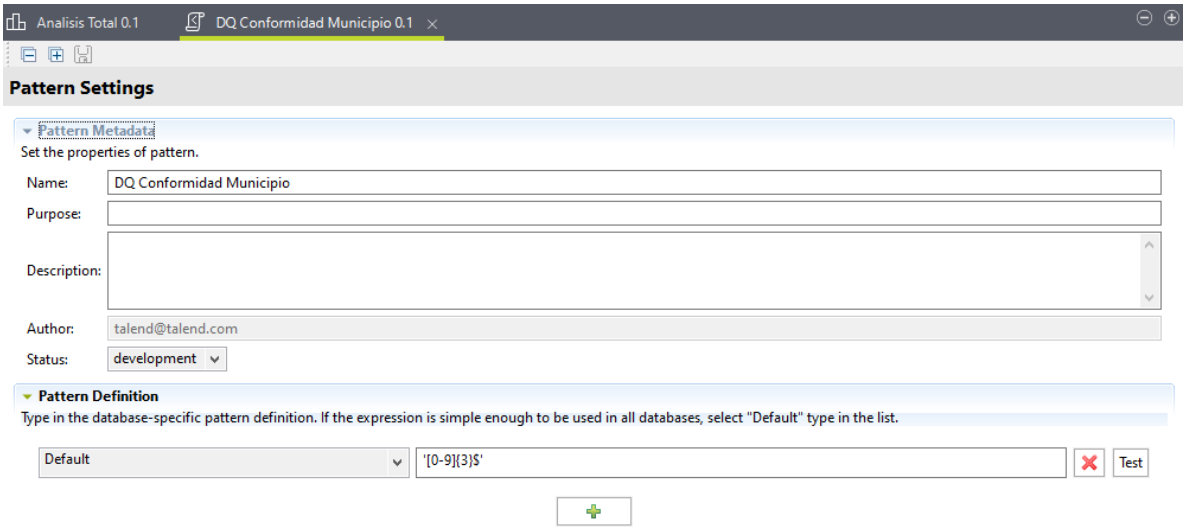


Ilustración 14: Conformidad municipio.

Validez Fechas

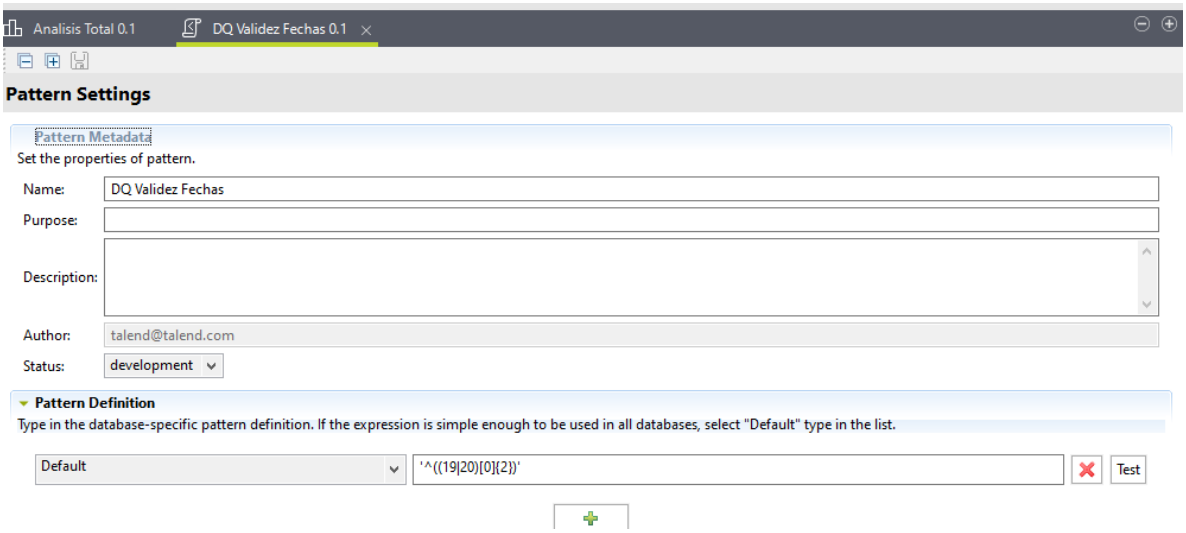


Ilustración 15: Validez Fechas

Valida Caracteres especiales.

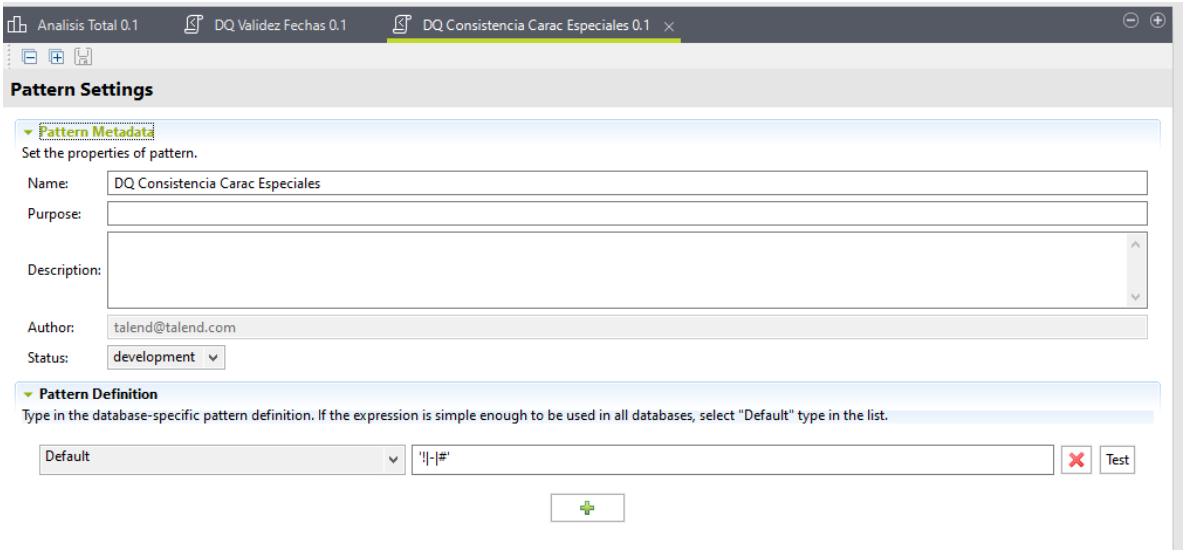


Ilustración 16: Valida Caracteres especiales

En este segmento se visualizan la columnas o elementos de datos críticos seleccionados:

Columnas perfiladas.

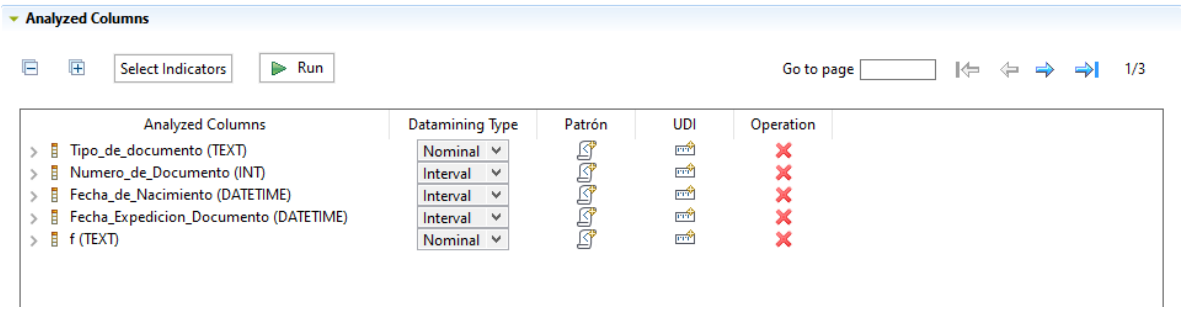


Ilustración 17: Columnas perfiladas.

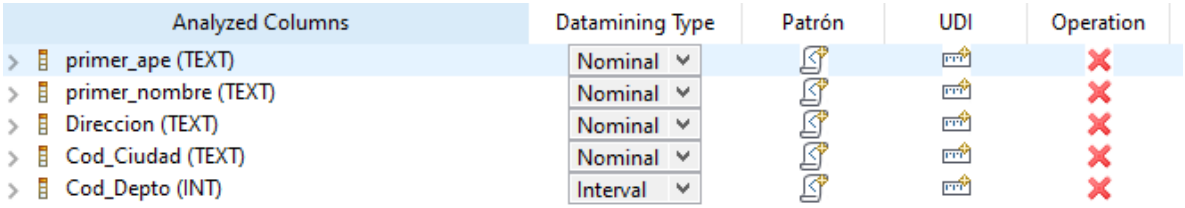


Ilustración 18: Columnas perfiladas.

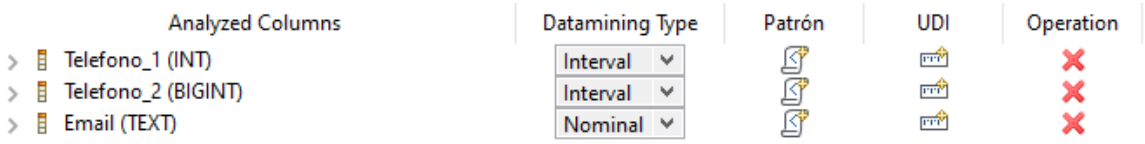


Ilustración 19: Columnas perfiladas.

Es importante mencionar que no sólo se utilizaron las reglas definidas por el negocio, sino que igualmente se hizo uso de las reglas para evaluar la calidad de los datos disponible por defecto en la herramienta:

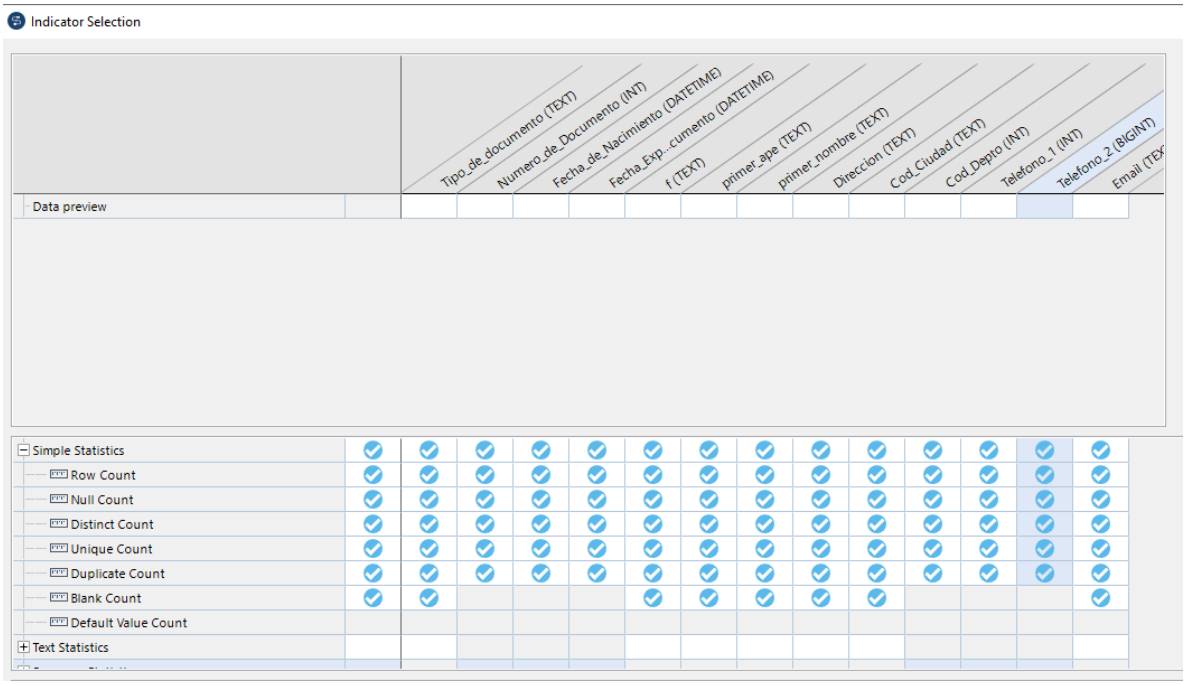


Ilustración 20: Reglas para evaluar la calidad de los datos precargadas en Talend Open Studio

La herramienta de calidad de datos Talend Open Studio permite a través de un panel configurar las reglas de acuerdo a las dimensiones aplicables:

Ilustración 21: Reglas diseñadas para evaluar la calidad de los datos en Talend Open Studio

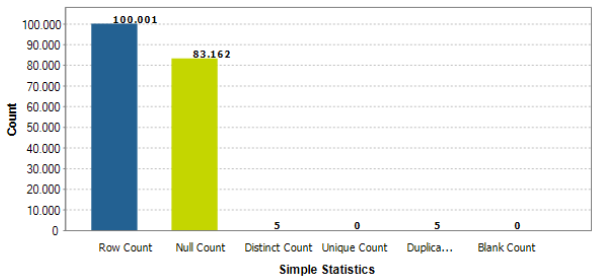
EL resultado de la aplicación de reglas se visualiza ejecutando el perfilado de los datos así:

Tipo de Documento.

Column: fact.Tipo_de_documento

Simple Statistics

Label	Count	%
Row Count	100001	100.00%
Null Count	83162	83.16%
Distinct Count	5	5E-3%
Unique Count	0	0.00%
Duplicate Count	5	5E-3%
Blank Count	0	0.00%



Pattern Matching

Label	Match%	Not Matc...	Match	Not Match
DQ Completitud Vacios	0.00%	100.00%	0	100001
DQ Consistencia Blancos	0.00%	100.00%	0	100001
DQ Consistencia Carac Especiales	0.00%	100.00%	0	100001

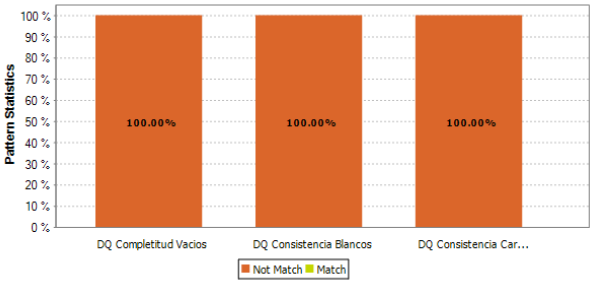


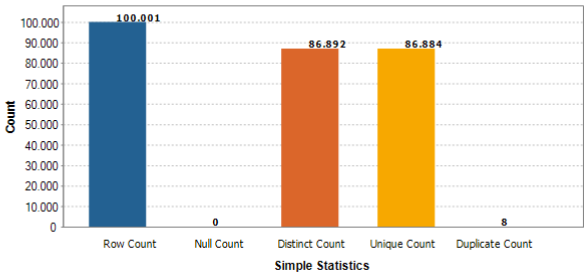
Ilustración 22: Tipo de Documento

Número de Documento.

Column: fact.Numero_de_Documento

Simple Statistics

Label	Count	%
Row Count	100001	100.00%
Null Count	0	0.00%
Distinct Count	86892	86.89%
Unique Count	86884	86.88%
Duplicate Count	8	8E-3%



Pattern Matching

Label	Match%	Not Matc...	Match	Not Match
DQ Duplicidad Numdoc	10.95%	89.05%	10947	89054
DQ Completitud Vacios	0.00%	100.00%	0	100001
DQ Consistencia Blancos	0.00%	100.00%	0	100001
DQ Consistencia Carac Especiales	0.00%	100.00%	0	100001

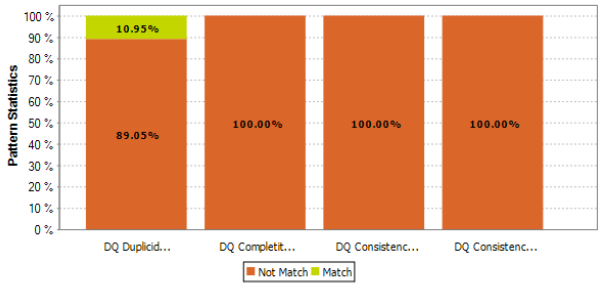


Ilustración 23: Número de Documento

Fecha de nacimiento.

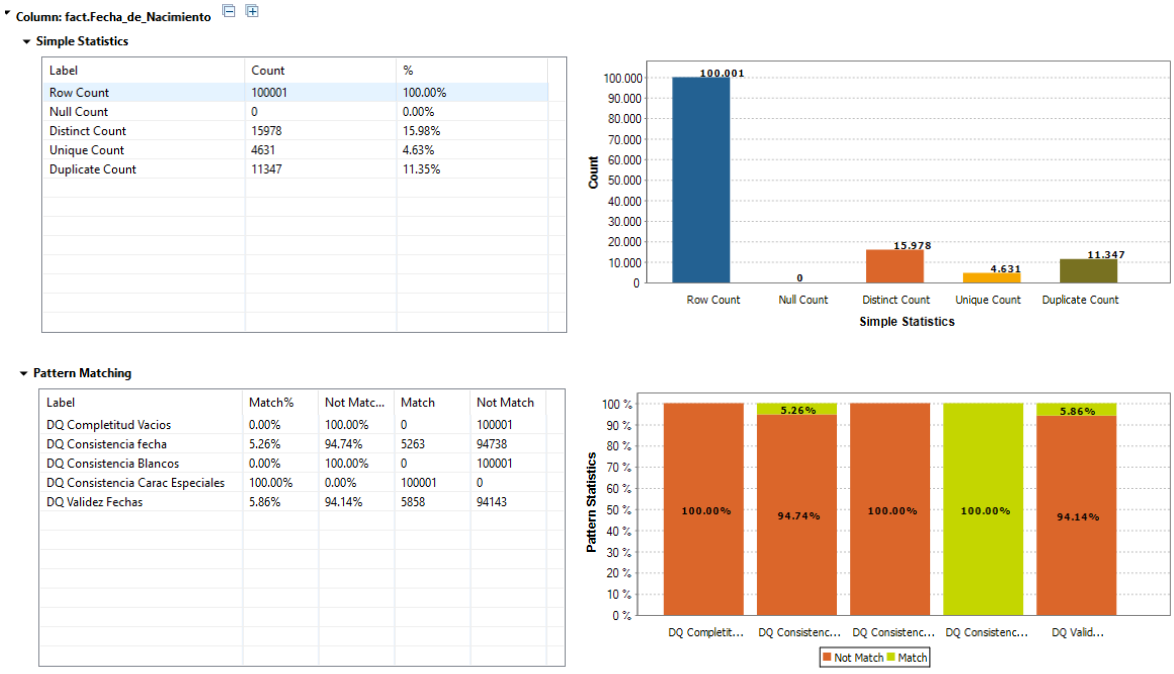


Ilustración 24: Fecha de nacimiento

Seguidamente se realiza el análisis de causa raíz utilizando la espina de pescado.

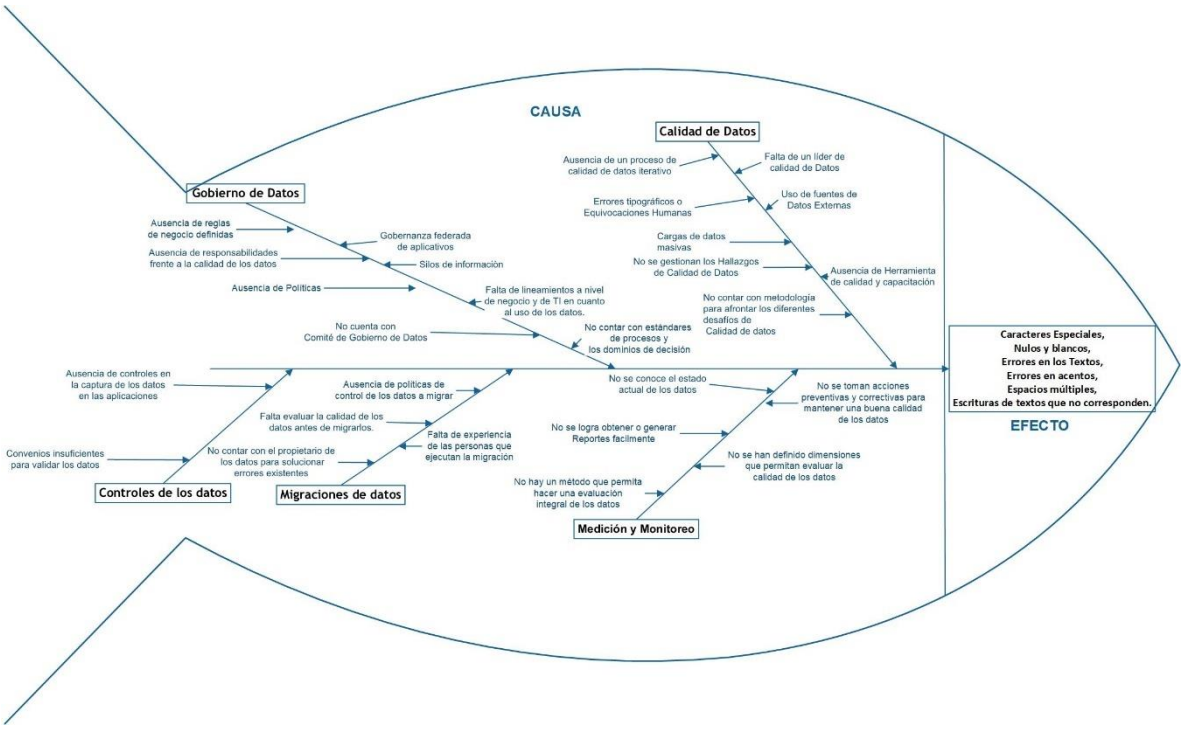


Ilustración 25: Análisis de causa raíz de problemas de datos en la organización

Teniendo en cuenta la ilustración anterior, el diagrama de espina de pescado consolida el análisis de las causas raíz de los problemas de calidad de datos evidenciados en el set de datos, esta actividad es necesaria realizarla con el negocio. A continuación, se listan las categorías de causa raíz:

1) **Gobierno de Datos**. En esta categoría se relacionan las causas raíz asociadas a la ausencia de políticas, procesos, lineamientos, roles y responsabilidades frente a la calidad de los datos.
2) **Calidad de Datos**. Las causas raíz que se listan en esta categoría son todas aquellas asociadas a la falta de una metodología rigurosa que permita mantener la alta calidad en los datos.
3) **Controles de los datos**. En esta categoría se mencionan las causas asociadas a la ausencia de controles en la captura de los datos en las aplicaciones.

4) **Migraciones de datos**. Las causas raíz que en esta categoría se mencionan son aquellas donde la organización no tiene control sobre cargas masivas y restricciones en campos obligatorios que generan mala calidad.
5) **Medición y monitoreo**. En esta categoría se identifican las causas raíz donde la organización no ha establecido las dimensiones o cualidades que deben tener los datos para ser evaluados, los métodos para determinar el estado integral de la calidad de los datos.

c) Medir y Monitorear la Calidad de los Datos

Para medir y monitorear la calidad de los datos, se inicia con priorización de las dimensiones de calidad de datos a utilizar para cada elemento de dato:
El cálculo de los pesos se realiza con ley de Borda-Kendall, teniendo en cuenta que se colocan la cantidad de dimensiones que se van a utilizar y de esta forma deben estar ordenadas de la más importante a la menos critica para el negocio.

POSICIÓN	Descripción de los criterios				
	Código	Dimensiones	Pesos	Dominio	Solución Ideal
1	C1	Completitud	0,67	[0,1]	[0.9 , 1]
2	C2	Consistencia	0,33	[0,1]	[0 , 0.05]
		Suma	1,00		

Tabla 6: Calculo peso dimensiones. Fuente propia

En el siguiente modulo se configura el umbral del ideal de referencia y los intervalos por cada una de las dimensiones y elementos de datos:

Ideal de Referencia	
-	-
0,03	0,03

Tabla 7: ideal de referencia. Fuente propia

Intervalos o Rango (Dominios de trabajo)	
0	0
1	1

Tabla 8: Intervalos o rangos. Fuente propia

Posteriormente se tabulan los resultados de cada una de las dimensiones aplicadas a cada elemento de dato:

Matriz de Decisión	
Completitud	Consistencia
0,83	0,1

Tabla 9: Matriz de decisión. Fuente propia

Matriz Normalizada	
Completitud	Consistencia
0,17526	0,92784

Tabla 10: Matriz normalizada. Fuente propia

Luego se consolidan todos los resultados:

Campos	Resultado	Evaluación	TOTAL
Tipo Documento	0,38	Baja Calidad	100.001
Numero Documento	0,92	Excelente	100.001
Fecha Nacimiento	0,70	Baja Calidad	100.001
Fecha Expedición	0,64	Baja Calidad	100.001
Primer Apellido	0,61	Baja Calidad	100.001
Primer Nombre	0,67	Baja Calidad	100.001
Direccion	0,69	Baja Calidad	100.001
Ciudad	0,90	Excelente	100.001
Departamento	0,63	Baja Calidad	100.001
Telefono1	0,80	Aceptable	100.001
Email	0,69	Baja Calidad	100.001

Telefono2	0,93	Excelente	100.001

Tabla 11: Resultados método RIM. Fuente propia

Como resultado de este ejercicio se presenta a la organización en la siguiente ilustración el Dahsboard de seguimiento integral a la calidad de los datos el cual muestra al negocio como están los datos de clientes a partir de la aplicación de la metodología que incluye las reglas de negocio para evaluar la calidad, dimensiones de calidad de datos, el método RIM, y los umbrales establecidos:

- Del 0 al 0.7 = Baja Calidad.
- Del 0.71 al 0.8 = Aceptable.
- Del 0.81 al 1 = Excelente.

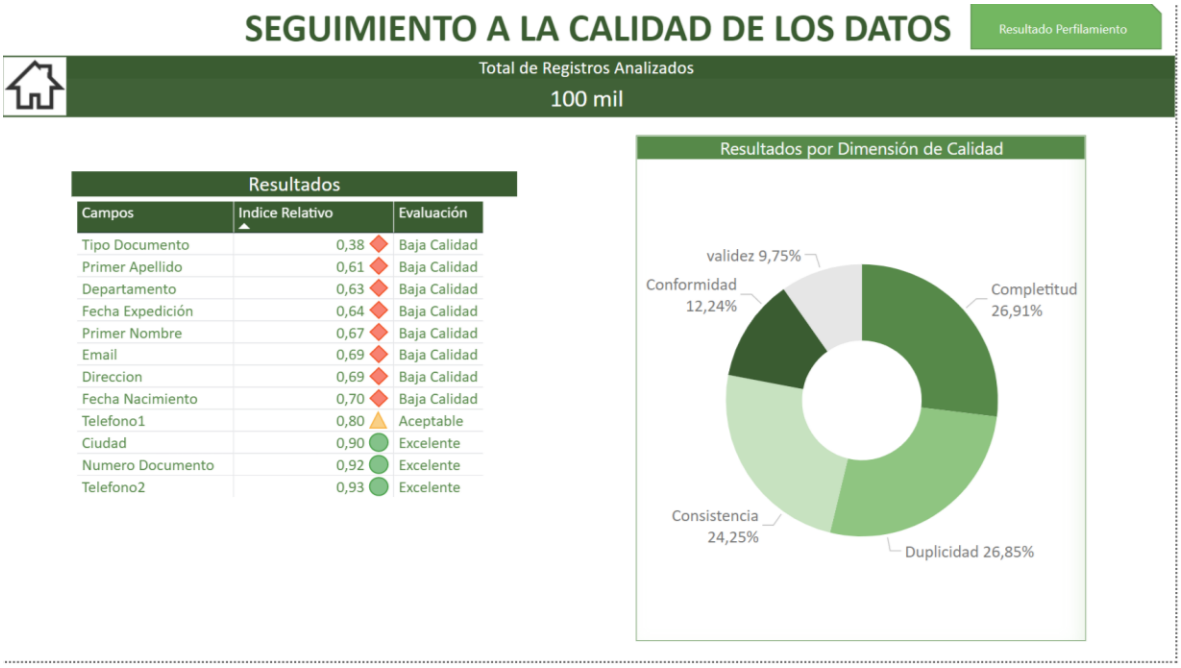


Ilustración 26: Resultado integral. Fuente propia

En resumen, la ilustración anterior presenta el Dashboard de seguimiento integral al estado de salud de los datos de una organización el cual se implementó con las siguientes funcionalidades:

- filtrar el resultado de cada uno de los elementos de datos.

- Filtrar por las dimensiones aplicadas a cada uno de los elementos de datos.
- A través del semáforo se puede comprender el nivel de la calidad de una forma integral por cada uno de los elementos de datos.

6.Conclusiones

Luego de evaluar algunos métodos para la evaluación de datos y de aplicar el método planteado podemos llegar a las siguientes conclusiones:

- El dashboard de seguimiento al estado actual de los datos en una organización se convierte en una herramienta clave para que las organizaciones logran tener una visión integral de las problemáticas que se presentan en los datos para tomar acciones que les permita mejorarlos y de esta manera impactar positivamente al negocio.
- Cabe resaltar que al revisar los diferentes métodos para la evaluación de la calidad se evidencia que no existe un cómo hacerlo, sin embargo, se pueden organizar los pasos para perder presentar una propuesta.
- Las organizaciones se pueden apoyar de un gran número de herramientas de calidad de datos que ofrecen funcionalidades que les van a facilitar las actividades de evaluación de la calidad, sin embargo, se deben revisar las necesidades puntuales de cada organización.
- La implementación del método multicriterio RIM, es relevante en su aplicación ya que genera un valor agregado porque permite entregar una evaluación integral de cada uno de los elementos de datos.

7.Referencias

1. Grupo PowerData. (Abr, 2017). Calidad de Datos. Cómo impulsar tu negocio con los datos. Powerdata.Es.https://www.powerdata.es/calidad-de-datos
2. Ibm.com. (May, 2019). Calidad de datos. https://www.ibm.com/co-es/analytics/data-quality
3. Iso8000.es. (Mar, 2018). Normas ISO 8000. http://iso8000.es/normas-iso-8000
4. iso.org (Sep, 2015). ISO 25012. https://iso25000.com/index.php/normas-iso-25000/iso-25012
5. iso.org. (oct 2020). ISO 22745. https://www.iso.org/standard/53995.html
6. Informatica.Com. (Oct, 2020). Mejore la calidad de sus datos para acelerar la transformación digital basada en datos. https://www.informatica.com/content/dam/informatica-com/es/collateral/data-sheet/es_data-quality_data-sheet_6710.pdf
7. Ibm.com. (n.d.). (Ago, 2021). DataStage - Visión general. https://www.ibm.com/co-es/products/datastage
8. Talend.com. (Jun, 2021). Los datos no saludables son un riesgo que no puede permitirse. https://www.talend.com/
9. Powerdata.es. (Mar, 2017). (n.d.). *Características y beneficios de Data Services SAP*. https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/caracteristicas-y-beneficios-de-data-services-sap
10. Gonzales, Hernandez. (Oct, 2017). *"diagnóstico de la calidad y el entendimiento de los datos para el análisis y toma de decisiones en las áreas de negocio de la empresa de telecomunicaciones xyzw*. https://alejandria.poligran.edu.co/bitstream/handle/10823/1127/TRABAJO%20DE%20GRADO.pdf?sequence=3&isAllowed=y

11. *E. Cables, M.T. Lamata, J.L. Verdegay (2016). "Reference ideal method in multicriteria decision making". Information Sciences, Vol. 337-338, pp. 1-10, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2015.12.011*.
12. Rodríguez, (2020), desarrollo RPA para monitoreo de calidad de datos y generación de alertas. https://repository.eafit.edu.co/bitstream/handle/10784/17555/Diego_RodriguezGarcia_2020.pdf?sequence=2&isAllowed=y
13. uexternado.edu.co. Rodríguez, (2019). *Plan de gestión de calidad de datos para mejorar la oportunidad y pertinencia de la información de la oferta institucional en la dirección de apropiación del ministerio tic magaly rincón rodríguez*. https://bdigital.uexternado.edu.co/bitstream/handle/001/2451/ABCBA-spa-2019-Plan_de_gestion_de_calidad_de_datos_para_mejorar_la_oportunidad_y_pertinencia?sequence=1&isAllowed=y
14. Beetrack.com. (Sep, 2020). *Ciclo de Deming: ejemplos, etapas, importancia, ventajas y desventajas*. https://www.beetrack.com/es/blog/ciclo-de-deming-etapas-ejemplos
15. geekflare.com. (October 11, 2021). *Las 5 mejores herramientas de gestión de datos para dar formato a sus datos para análisis*. https://geekflare.com/es/best-data-wrangling-tools/
16. Zipforecasting.com. (Oct, 2020). *Evaluacion de la calidad de los datos-Metricas y pasos para conocer*. https://zipforecasting.com/es/data-driven/data-quality-assessment.html
17. Powerdata.Es. (Abr, 2017). *Calidad de Datos. Cómo impulsar tu negocio con los datos*. https://www.powerdata.es/calidad-de-datos

18. Coleman. (dic 31, 2013). Measuring Data Quality for Ongoing Improvement. https://www.elsevier.com/books/measuring-data-quality-for-ongoing-improvement/sebastian-coleman/978-0-12-397033-6

19. McGilvray. (Jul 25, 2008). Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information. https://www.pdfdrive.com/executing-data-quality-projects-ten-steps-to-quality-data-and-trusted-information-tm-e158581809.html

20. Strong, Wang. (1997). 10 Potholes in the Road to Information Quality. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/http://web.mit.edu/tdqm/www/tdqmpub/10potholesIEEEComputerAug97.pdf

21. **Shewhart. (1931)**. Control económico de la calidad de productos manufacturados. https://www.editdiazdesantos.com/libros/shewhart-wa-control-economico-de-la-calidad-de-productos-manufacturados-L03003040101.html

22. **Crosby. (1987).** La calidad no cuesta. https://www.academia.edu/8118377/La_calidad_no_cuesta_Philip_B_Crosby