Análisis del proceso de calidad de datos y estudios comparativos de herramientas Open Source sobre perfilado de datos, enfocado a la mediana y pequeña empresa.

Natalia Andrea Jaramillo

Mile Yurley Orrego Porras

Carlos Eduardo Ossa Quintero

Ingeniería de sistemas

Director(es) del trabajo de grado

Gustavo Macias Suárez

INSTITUTO TECNOLÓGICO METROPOLITANO

2017 24 11

RESUMEN

Desde la Inteligencia de negocios se analiza la calidad de los datos, es así como en los diferentes procesos que se llevan a cabo en las medianas o pequeñas empresas se utiliza información, desde el mismo momento de su recolección, el procesamiento y el almacenamiento se debe contemplar la calidad e integridad de la misma. El objetivo del presente estado del arte es realizar una traza de cinco aspectos significativos a tener en cuenta en el momento de realizar el proceso de la calidad del dato, como son: gestión, dimensión, perfilamiento, enriquecimiento y transformación. Se tiene en cuenta también, la perspectiva de la gestión de la información en los datos capturados, procesados, almacenados y entregados al usuario; el cual debe ser un fiel reflejo de la realidad que se desea tratar con los sistemas informáticos, generalmente de administración como son los SGBD. La metodología aplicada está enmarcada en el análisis de tres herramientas que utilizan la calidad de datos como son: SQL Power DQguru, Talend Open Profiler y Google Refine a partir de una o varias bases de datos que están normalizadas.

Palabras clave: Calidad de datos, proceso de calidad, herramientas de Limpieza y perfilado.

RECONOCIMIENTOS

Agradecimientos a todas aquellas personas que con su ayuda han colaborado en la realización del presente trabajo, en especial al profesor Jorge Iván Bedoya, el asesor del trabajo de grado, por la orientación, el seguimiento y la supervisión continúa de la misma, pero sobre todo por la motivación y el apoyo en el semillero de BI. Por último, un agradecimiento profundo a nuestros padres quienes nos han apoyado en todo el proceso de la carrera.

ACRÓNIMOS

BD: Base de datos es un “Almacén” que nos permite guardar grandes cantidades de información. (Valdés, 2007).

Data Cleaning: Limpieza de datos es el acto de descubrimiento *y* correcciones o eliminación de registros de datos.

CVS: (del inglés comma-separated values) Datos separados por comas.

TXT: Extensión de archivo de texto.

XML: Es un subconjunto de SGML(Estándar Generalised Mark-up Language),simplificado y adaptado a Internet.

JSON: Los archivos JSON se utilizan para almacenar estructuras de datos simples utilizando un formato basado en texto legible.

XLS: Formato de archivo de Microsoft Excel.

SQL: (Structured Query Language) es un lenguaje de programación estándar e interactivo para la obtención de información desde una base de datos y para actualizarla. (Rouse, s.f.)

BSD: son las siglas de “Berkeley Software Distribution”.

GREL: Google Refine Expression Language, es el leguaje de expresiones regulares que tiene Open Refine.

JDBC: Java Database Connectivity, driver que permite interactuar con varias bases de datos.

Pymes: Acrónimo de empresas pequeñas y medianas. Se trata de la empresa mercantil, industrial o de otro tipo que tiene un número reducido de trabajadores y que registra ingresos moderados (Gardey, 2009)

TABLA DE CONTENIDO

INDICE DE TABLAS E ILUSTRACIONES

1. INTRODUCCIÓN

Cada vez está tomando mayor fuerza la calidad de datos, a partir del impacto que está generando en las empresas la pérdida de información. La calidad, según la real academia de la lengua, es la propiedad o conjunto de propiedades inherentes a algo, que nos permite juzgar su valor. Hora planteamos ¿qué es dato? "Son un término general para denotar alguno o todos los hechos, letras, símbolos y números referidos, o que describen, idea, situación, condición u otro factor" (J.Maynard-Smith, 1982). Este trabajo de grado tiene como objetivo, estudiar tres herramientas para el perfilamiento de datos derivadas de gestión y dimensión de la información.

Generalidades

En la actualidad las empresas tienen un sin número de datos que pueden tener valor o no, la oportunidad que genera estas es darle valor a esos datos que no tienen relevancia, un tratamiento adecuado de la información significa generar oportunidad de negocios para que las empresas crezcan en la medida del tiempo y fortalecer así su calidad de datos.

Justificación

El propósito central de este trabajo de grado es analizar la calidad de los datos que las PYMES tienen. A estos datos se les da un tratamiento teniendo en cuenta la estructura como están integrados y por ende se realiza su gestión, dimensión, perfilamiento, enriquecimiento y transformación, por medio de tres herramientas open source que tienen diferentes características, pero ejercen la misma filosofía que es dar una calidad a los datos, con el fin de saber que algunas empresas tomen estas como un medio para realizar una buena inteligencia de negocios.

En la ejecución del análisis habrá fases que determinaran qué mejoras se le pueden hacer a los datos y así encontrar qué delimitantes tiene el trabajo de grado. La primera de ellas es obtener información sobre el proceso de calidad y perfilamiento de los datos para plasmarlo en un marco teórico; en la segunda parte de este proyecto se analizará la base de datos que se facilitó para la ejecución; por último, en la tercera parte se ejecutará el análisis realizado donde se mostrará qué cualidades tiene cada una, así dar un insumo a las empresas sobre la herramienta que potencialmente le favorece más dentro de su mercado y para la toma de decisiones.

Planteamiento del problema

En las PYMES se vienen presentando dificultades en el manejo de la calidad de los datos y no se ven con claridad las respuestas a los cuestionamientos que según la norma ISO 25012 propone a las mismas sobre ¿Qué políticas de calidad de datos o normas ISO deberían establecerse en cada compañía y cuál de las herramientas de Open Source que se eligieron es la más efectiva para llevar a cabo el cumplimiento de dichas políticas? La pregunta planteada busca la relación entre los siguientes dos aspectos:

1. El conocimiento sobre calidad de datos en beneficio de las medianas y pequeñas empresas.
2. Qué utilidad de oportunidad tiene las herramientas de calidad de datos Open Source en PYMES, para que éstas tengan una idea del negocio enfocada a la toma de buenas decisiones como principal ventaja competitiva.

Objetivos

1.1 General

Realizar un estudio comparativo de tres herramientas Open Source sobre el perfilado de los datos, teniendo en cuenta la gestión, dimensión, perfilamiento, enriquecimiento, transformación de los datos, para realizar métricas definidas, con el fin de darle insumos a las PYMES y así decidir que herramientas a utilizar.

1.2 Específicos

- Efectuar el estado del arte a partir del estudio sobre el perfilamiento y calidad de los datos.
- Plasmar un estudio de las herramientas Open Source para el perfilado y calidad de datos.
- Analizar el nivel de los datos entregados, para así determinar las mejores herramientas a emplear en un perfilamiento de datos.
- Realizar un enfoque desde la gestión, dimensión, perfilamiento, enriquecimiento, transformación de los datos que permita tener una mayor claridad sobre los resultados que se requieren obtener.
- Realizar un análisis sobre la calidad de los datos que permita descubrir las principales anomalías, enfocados hacia los datos maestros para construir la toma de decisiones.

Organización de la tesis.

La primera parte del trabajo contiene el marco teorio sobre la calidad de los datos ¿qué es?, su dimencionamiento, gestion y la tranformacion de los datos.

En la segunda parte se trata de la metodología que contiene la estrategía con la cual se hará el análisis y experimentación de las herramientas de perfilamiento de datos que es objeto de estudio.

Por último, la tercera parte contiene la ejecución del análisis y experimentación sobre las herramientas y los resultados que son arrojados de este.

1. MARCO TEÓRICO

Para determinar el concepto de calidad de datos, primero se debe plantear el significado de calidad, la calidad, en su descripción de la real academia de la lengua, es la propiedad o conjunto de propiedades inherentes a algo, que nos permite juzgar su valor. Ahora Planteemos ¿Qué es dato?, los datos son "un término general para denotar alguno o todos los hechos, letras, símbolos y números referidos a, o que describen, idea, situación, condición u otro factor" y que constituyen un elemento fundamental para la toma de decisiones objetivas a todos los niveles de una organización (Vilalta Alfonso, 2018)

Conociendo los conceptos de cada uno de los términos, la calidad de los datos "examina si los datos de una organización son confiables, consistentes, actualizados, están libres de duplicidades y si son apropiados para sus objetivos". (Vilalta Alfonso, 2018); o si bien la calidad de los datos son un conjunto de propiedades inherentes al dato que permite determinar si el mismo es correcto o incorrecto. Estas propiedades se denominan dimensiones de la calidad del dato, algunas de ellas son: eficacia y veracidad.

Dimensión de la calidad de datos.

Según Jhon A Hoxmeier en su artículo: A Framework for Assessing DataBase Quality, "las dimensiones de una calidad de bases de datos son principalmente procesos y los datos; sin embargo, se ha realizado estudios sobre nuevas técnicas para el modelaje de las bases de datos, las cuales incorporan nuevas dimensiones al estudio de la calidad de las mismas, como son: la semántica y el comportamiento. (S. Wilke, 2014)

Las dimensiones de los datos están constituidas por: antigüedad del dato o periodo de tiempo en que el mismo no ha cambiado; las características propias del dato, tales como la exactitud en las diferentes fuentes donde éstas se encuentran almacenadas; el contexto dado, que es representado por el contenido o valores que puede adquirir: seguridad, representa los diferentes roles que se definen para acceder a los datos; y por último, el modelaje de datos, que está constituido por la flexibilidad, el contenido, el alcance, su normalización y relevancia.

La metodología para el diagnóstico de la calidad de los datos tiene procedimientos para los diagnósticos de esté y se cumplen en diferentes etapas tal como lo menciona : (Vilalta Alfonso, 2018)

- Identificación y ordenamiento de los datos críticos de clientes.
- Identificación a partir de los clientes

- Identificación a partir de los especialistas
- Selección del tipo de datos crítico a diagnosticar según el listado ordenado.
- Definir un responsable de calidad del dato objeto de estudio.
- Analizar las dimensiones de calidad del dato.
- Determinar las dimensiones de calidad asociadas al dato.
- Identificar los indicadores asociados a las dimensiones de calidad del dato y determinación de la frecuencia de medición de cada indicador.
- Realizar las mediciones necesarias para analizar las dimensiones de calidad del dato.
- Localizar la información asociada al dato y toma de muestras.
- Calcular los indicadores y detectar problemas
- Procesamiento de los indicadores de calidad asociados a los datos.
- Elegir las herramientas apropiadas para el análisis de los indicadores.
- Desarrollo de la(s) herramienta(s) seleccionada(s).
- Análisis de los problemas de calidad del dato.
- Diagnóstico de la calidad del dato.
- Monitoreo de los indicadores por parte de la gerencia.
- Garantizar la seguridad del dato.
- Determinar el nivel de seguridad en que se encuentra el dato.
- Elaborar el manual de seguridad del dato.

Por último, analizar la etapa del dimensionamiento de calidad de dato, a lo cual se pretende establecer criterios (Vilalta Alfonso, 2018) que permitan garantizar la calidad del dato, o sea, permiten garantizar que el dato sea exacto, integro, consistente, puntual, exclusivo, valido. Además, una vez determinadas las dimensiones, se procede a localizar los indicadores o parámetros, ya sean cuantitativos o cualitativos, que representen y garanticen el cumplimiento de las dimensiones de calidad asociada al dato.

Sistema de gestión de datos.

Para el sistema de gestión Schoenbach (2004) Afirma que:

> El sistema de gestión de datos es el conjunto de procedimientos y personas por medio de los cuales se procesa la información. Involucra la recolección, manipulación, almacenamiento, y recuperación de información.
>
> Como tal, se espera obtener un medio que permita procesar adecuadamente la información sin que vea mayormente afectada por la manipulación que sea hace desde la recolección de los datos hasta la muestra final de los mismos. Para ello cabe resaltar los objetivos que el sistema de gestión debe cumplir y que se citaran a continuación:
>
> El objetivo del sistema de gestión de datos es el de asegurar:
>
> a) Datos de alta calidad, i.e., asegurar que las variabilidades en los datos provienen del fenómeno en estudio y no del proceso de recolección de datos.

b) Un análisis e interpretación de datos precisos, apropiados y defendibles.

Transformación de los datos

"Después de la fase de exploración, el proceso de extracción del conocimiento contempla la fase de limpieza de datos (data cleaning)." (Lopez, 2007). Mediante dicho proceso es necesario realizar el ajuste o creación de algoritmos más robustos que suplan a tal medida la necesidad de realizar un filtro más preciso de la información, ya que se tienes valores atípicos, faltantes o erróneos.

Para ello se lleva a cabo una transformación de los datos con diferentes técnicas que permitan aumentar, reducir o modificar por completo la dimensión de los datos. Entre las técnicas más avanzadas se encuentran la reducción y aumento de la dimensión.

Independientemente de su procedencia, los datos son sometidos a procesos o técnicas diferentes con el objetivo de perfilarlos y homogeneizarlos para que puedan cumplir su función de generadores de información de calidad útil y efectiva para la toma de decisiones.

Como lo indica López (2008) para las transformaciones se consideran cuatro tipos:

- Transformaciones lógicas: "Se unen categorías de campo de definición de las variables para reducir así su amplitud.", con lo cual se permite establecer una relación de orden de objetos, números, distancias, entre otros.
- Transformaciones lineales: "Se obtienen al sumar, restar, multiplicar o dividir las observaciones originales por una constante para mejorar su interpretación.", al transformar datos numéricos se pretende tener una mayor fiabilidad de los datos para una correcta elección al momento de tomar decisiones.
- Transformaciones algebraicas: "Se obtienen al aplicar transformaciones no lineales monotónicas a las observaciones originales (raíz cuadrada, logaritmos, etc.) por una constante para mejorar su interpretación.", analizar la cantidades de operaciones algebraicas que se aplican en cada una de las transformaciones para poder expresar compacta y eficiente las diferencias clases de objetos matemáticos.
- Transformaciones no lineales no monotónicas: "Cambian las distancias y el orden entre los valores.", teniendo funciones entre conjuntos ordenados que conservan el orden dado.

La transformación es posible gracias a la necesidad de arreglar problemas en los datos, lo cual permite que una vez sean transformados, los datos medidos en escalas diferentes sean más comparables entre sí, que dentro del marco de inteligencia de negocios amplia el asertividad de la transformación de los datos en decisiones como una estrategia empresarial. Para este entorno de negocios que cada vez es más competitivo y desafiante, las empresas tienen la necesidad de encontrar soluciones y sistemas para generar ventajas competitivas a partir de la recopilación, análisis y transformación de datos en decisiones estratégicas que les permitan diseñar planes exitosos y gestionar adecuadamente las distintas áreas y departamentos.

Enriquecimiento de los datos

El enriquecimiento de datos es un proceso que complementa y enriquece la información en una base de datos existente, que permitirá tomar mejores decisiones. Este proceso se engloba en el data mining. “El data mining (minería de datos), es el conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto.” (Sinnexus, s.f.) Es posible que, para aplicar un enriquecimiento de los datos, se aplique métodos como:

- Análisis: Es realizado para la detección de errores de sintaxis.
- Transformación de Datos: La transformación de datos permite al trazar un mapa de datos, en el formato esperado. Esto incluye conversiones de valor o funciones de traducción, así como normalización de valores numéricos.
- Eliminación de duplicados: La detección de duplicados requiere un algoritmo para determinar si los datos contienen representaciones dobles de la misma entidad, siendo así más fácil ser comparables.
- Método Estadístico: Incluye analizar los datos usando promedios, desviación estándar, rangos, o algoritmos de cluster, este análisis se realiza por expertos que identificar errores. Aunque la corrección de datos sea difícil ya que no saben el valor verdadero, pueden ser resueltos poniendo los valores a un promedio u otro valor estadístico.

Perfilamiento datos

La información que proporcionan los perfiles de datos es de gran ayuda para mejorar la calidad de los datos. En concreto, el proceso de perfilado evalúa y mejora la calidad de los datos de un sistema de origen dado, tratando de corregir y mejorar los problemas existentes, así como evitarlos en el futuro.

Los proyectos de calidad de datos pueden contar con soluciones de perfilado de datos que automaticen el proceso que suponen una evolución cualitativa con respecto al perfilado de codificación manual. Entre otras ventajas, reducen el esfuerzo y amplían el alcance y la mejora de la coherencia en todas las iniciativas de calidad de datos. Al contar con capacidades de descubrimiento automatizadas escanean todos los registros de datos únicos, de cualquier fuente, encontrar anomalías y relaciones ocultas con gran eficiencia. (PowerData, 2016)

Herramientas de para estudio comparativo

SQL Power DQguru

La herramienta de SQL Power DQGuro es predilecta ya que está integrada con varios SGBD y permite realizar un modelo de perfilado y calidad de datos, siendo en esencia “ideal para la limpieza de

cualquier Data Warehouse o base de datos CRM, permite realizar procesos de limpieza de datos y gestión de datos maestros, identificando y eliminando los duplicados, construyendo referencias cruzadas entre las tablas de origen y destino. Esto les proporciona a los usuarios los datos completos y precisos, una sola visión de 360 grados de todas las entidades de negocio. Esta es una herramienta de Data Cleansing que SQLPower ha liberado convirtiendo la licencia en Open Source. Su funcionamiento es sencillo, consta de crear un repositorio sobre una de las diferentes bases de datos a trabajar con las que conecta por JDBC, y se pueden crear proyectos de 3 tipos diferentes: Deduplicación, Datacleansing y Referencias cruzadas. La interfaz para realizar estas acciones es muy intuitiva y visual. Aunque hay operadores como los de comparación fonética, se echan de menos funciones de fuzzy logic para comparar palabras parecidas, o que se trabaje un porcentaje de similitud por campo y por registro. La herramienta muestra de una manera muy visual las coincidencias encontradas, con un color para cada proceso definido, y permite ver las diferencias entre registros, y descartar coincidencias, decidir cuál es el registro maestro (el que va a conservar los datos tras la fusión), y qué es lo que se va a fusionar y cómo. Finalmente, en el proceso de fusión deja un log y guarda los identificadores de lo que se fusiona en una tabla de resultados. Trabaja directamente sobre la tabla origen, y borra los registros que se han marcado como duplicados." (SQLPower, 2017)

Talend Open Profiler

Herramienta ideal que permite navegar por los esquemas de tablas de una base de datos para realizar una serie de análisis sobre la información de dependencias, número de registros, índices, valores nulos, longitud mínima o máxima, valores duplicados y demás datos que permiten identificar patrones en validaciones y verificaciones de la información. Con el análisis de la información, la herramienta genera gráficas de resultados que equivalen a los análisis de campos o patrones predefinidos que establecen una visualización más amplia para el trato de la información, de tal manera que tome medidas de corrección sobre ellos y tenga en cuenta en la definición de procesos ETL. De manera comercial “Se trata de una herramienta para gestionar los datos de candidatos a compañías. Los clientes se benefician de una solución de gestión de talento que se integra completamente con sistemas de recursos humanos básicos de registro, y con el mayor ecosistema de ERP. Los módulos de gestión del talento cubren todas las fases del ciclo del talento, como lo son la planificación, reclutamiento, el rendimiento, el aprendizaje, el desarrollo profesional, compensación, opiniones sobre el talento, medición y presentación de informes. Una de las características principales sobre la herramienta es la facilidad con la que el usuario puede hacer un adecuado uso de los datos adquiridos. Mediante uno de los procesos de data cleansing que se emplean, permite limpiar, validar y corregir datos de las fuentes de origen durante el proceso de incorporación y creación del Data Warehuose.” (Talend, 2017)

Google Refine.

La aplicación google refine es una herramienta eficiente la cual nos ayuda a organizar, permitiendo limpiar y transformar los datos alterados**.** No es un servicio web sino una aplicación de escritorio, esto permitirá que los datos estén seguros, aunque interactúe vía web. (Valles, 2013), actualmente esta herramienta ha cambiado de nombre por Open Refine pero no cambia su filosofía actual el cual "nos permite limpiar bases de datos, exportarlas en diferentes formatos, y arreglar y manejar las bases para un mejor uso. Actualmente el proyecto ya no es financiado por Google y se encuentra como proyecto abierto. Los archivos que podemos importar para trabajar pueden tener las extensiones TSV, CSV, XML, JSON, XLS, e incluso Google Spreadsheets, entre otros. También nos permite transformar archivos de cualquiera de estos formatos a otro. Open Refine funciona como ejecutable sobre cualquier navegador web y está disponible para Windows, Mac y Linux." (Rios, 2014)

2. METODOLOGÍA

Se pretende alcanzar los objetivos del análisis y experimentación de cada una de las herramientas propuestas anteriormente desde el estado del arte. Se planea una fase inicial de tres meses, para realizar el estudio de estas herramientas plenamente identificadas en el cronograma de actividades, en este periodo estará todo el análisis y las conclusiones propuestas.

Los datos entregados para el análisis y experimentación vienen de una base de datos llamada Todoterreno que es gratuita y ofrecida desde internet por la empresa ofimática S.A, cuyos datos son adecuados para realizar el proceso de calidad de datos, acto seguido se analizará esté, para saber qué datos se pueden realizar para dicho proceso, además de identificar características y consistencia para efectuar la transformación de los datos con cada una de las herramientas.

Los datos encontrados y adecuados para el análisis, se realizará la gestión, dimensión, perfilamiento, enriquecimiento y transformación de los datos a la cuales se le realizará procedimiento para obtener consistencia y cardinalidad.

Cada integrante tiene una herramienta asignada con la cual se iniciará un perfilamiento con la base de datos Todoterreno. Mediante la realización del proceso de perfilamiento de datos con cada una de las herramientas, se aplicarán reglas o expresiones regulares que permitirá la transformación de los datos, esperando un resultado igual o variable según los siguientes criterios:

1. Exactitud: concerniente a que tan preciso son los datos presentados, esperando que el resultado obtenido asegure la fidelidad de los datos en base a las reglas o expresiones regulares aplicadas en cada una de las herramientas presentadas.
2. Objetividad: Basada en datos comprobables que concedan la aceptación o no de la información.
3. Cobertura: Enfocada en que tan precisa es la aplicación de las reglas o expresiones regulares sobre determinada cantidad de datos.
4. Duplicación: Es importante tener en cuenta que si la información está en formatos iguales o parejos en la tabla que se esté analizando para posteriormente realizar el proceso con cada una de las herramientas.

Adicional a este análisis en la comparación de las herramientas se realizará la matriz de riesgo cualitativa, por la cual se obtendrá una visión complementaria de cada herramienta, permitiendo que desde este punto de vista se pueda llegar a concluir si pueden ser o no altamente efectivas durante un proceso de calidad de datos.

De acuerdo a lo anterior, cada uno usará para el perfilado de los datos una copia de la base de datos Todoterreno, de la cual se espera que durante la experimentación se obtengan resultado asertivo y

al finalizar se pueda emplear la matriz de riesgos para poder terminar de comparar todas las herramientas. Para la distribución en el análisis y experimentación con cada una de las herramientas para el proceso de calidad de datos se realizará de la siguiente manera:

- La ingeniera (c) Mile Yurley Orrego Porras tiene asignada a la herramienta Taled Open Profiler.
- La ingeniera (c) Natalia Andrea Jaramillo tiene asignada a la herramienta Power SQL Guru.
- El ingeniero (c) Carlos Eduardo Ossa Quintero tiene asignada a la herramienta Google Refine

Experimentación

La metodología implementada parte de la recolección, procesamiento y almacenamiento de la información, además de tener un conocimiento previo de la gestión, dimensión, perfilamiento, enriquecimiento y transformación de los datos, para posteriormente examinar los datos entregados con las herramientas: SQL Power DQguru, Talend Open Profiler y Google Refine.

Para la implementación de las pruebas y análisis de las herramientas de perfilamiento de datos es indispensable conocer las especificaciones a las cuales se ata el uso de las mismas. A continuación, se presentan los requisitos mínimos que se tienen en cuenta por cada una de las herramientas utilizadas:

Herramientas / Requisitos	SQL Power DQguru	Talend Open Profiler	Google Refine
Licencia	No tiene licencia, es una herramienta open source.	No tiene licencia, es una herramienta open source.	BSD
Sistema Operativo	Cualquier SO que soporte la versión 6.0 o superior de JRE	- OSX -Solaris -Ubuntu Linux -Microsoft Windows	-Ubuntu Linux -Microsoft Windows - OSX
Procesador	Procesador dual core 2.0 Ghz o superior	Procesador dual core 2.0 Ghz o superior	Procesador dual core 2.0 Ghz o superior
Memoria	2GB de RAM	2GB de RAM	2GB de RAM
Almacenamiento	2GB o superior de espacio disponible	2GB o superior de espacio disponible	2GB o superior de espacio disponible

Tabla 1 Fuente (SQLPower Software, 2018), (Talend, 2017), (Stephens, 2017)

Como material de insumo se tiene una base de datos que son de acceso restringido, sin embargo sirvieron de apoyo para las pruebas que se realizaron con las herramientas a las cuales se está evaluando, cabe acotar que las BD están normalizadas lo que permite aplicar reglas para obtener datos más organizados o limpios, donde se incluye la creación de tablas y el establecimiento de

relaciones entre ellas según reglas diseñadas tanto para proteger los datos como para hacer que la base de datos sea más flexible al eliminar la redundancia y las dependencias incoherentes.

Para establecer la conexión con diferentes bases de datos es necesario destinar un driver JDBC que permite la interacción adecuada con las aplicaciones SQL DQGuru y Talend Open Profiler, por lo contrario, las herramientas SQL Refine no requiere este complemento.

Después de la configuración de cada una de las herramientas se procede a experimentar con la base de datos:

a. SQL Power DQ Guru: En la cual se procede a realizar el diagrama de calidad de datos de esta herramienta, donde facilita la unión de las tablas, eligiendo las columnas según la necesidad del proceso, concatenándose según los parámetros de estandarización o expresión regular que requiera la columna de la tabla.

b. Talend Open Profiler: Se compone de dos diferentes sub-herramientas que realizan un proceso dependiente dirigido a un mismo fin, para determinar cuál es la mejor forma de optimizar los datos del BD. La primera herramienta Data Profiler realiza todo el análisis de las estructuras de las tablas arrojando una serie de estadísticas, donde se determina que columnas son las adecuadas para realizar todo el proceso de calidad de datos. Posteriormente a partir de los resultados obtenidos del anterior análisis, la segunda herramienta Data Quality realiza un proceso similiar al Power DQ Guru, concatenando cualquier tipo de dato donde se analiza su estructura, luego se aplica las reglas de expresión regular y se definen los valores a reemplazar, por último, ejecutando al nuevo cambio.

c. Google Refine: Básicamente hay que extraer los datos a refinar desde la BD que se tiene para llevarlos a un archivo plano pueda ser en TXT, CSV además de tener otros formatos admitidos, después de realizar este paso, se define que columnas pueden ser candidatas para realizar la limpieza, además de tener ciertas herramientas para realizar este, permite hallar erratas e inconsistencias de forma automática, empleando como se conoce “Clustering” en el cual consiste detectar asociaciones de valores muy parejos entre sí.

En esta fase de la experimentación es importante tener en cuenta que el modelo de la base de datos y mostrar las entidades y relaciones que tiene las tablas a intervenir en este caso se mostrara el siguiente modelo:

Diagrama de realaciones, Catalogo de terceros contables.

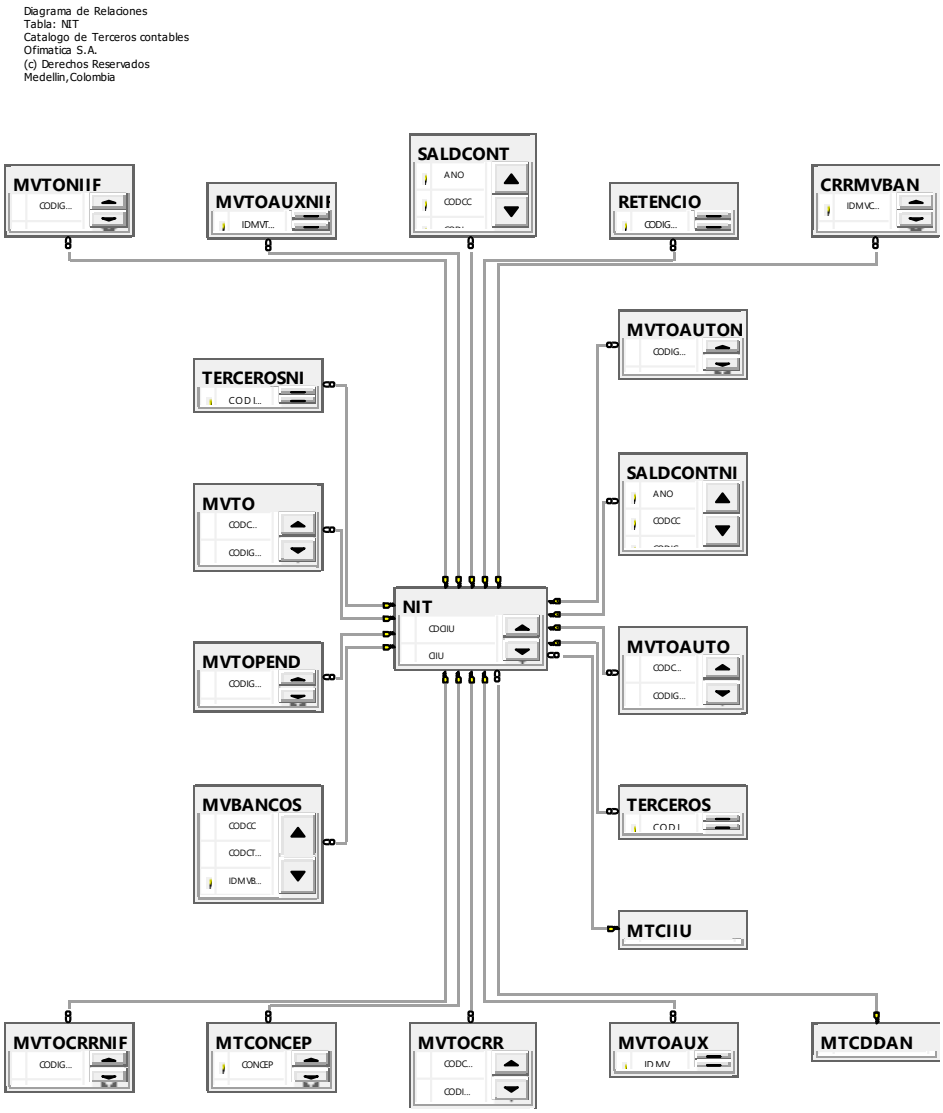


Ilustración 1 Relaciones de Catalogos Terceros fuente Ofimatica SA

Tabla 2

Propiedades de la tabla NIT

Nombre	NIT
Tamaño	93 Registros
Columnas	34
Tipo	Tabla de hechos
Uso	Transaccional
Versión DB	SQL Server

Tabla 2 NIT fuente DataBase Ofimatica S.A

Dado un universo de datos, se contempló la posibilidad de analizar el universo por completo de los datos, ya que la cantidad total de registros resultaba perfectamente manejable, para cumplir con los objetivos propuestos.

Del modelo de la BD, la tabla NIT se analizará el cambio y debido a este afectará a todos los datos que este contiene, específicamente se aplicará los cambios a la columna dirección como muestra la siguiente imagen:

DIRECCION
118400 MAIN STREET, LOS ANGELES
118400 MAIN STREET, LOS ANGELES
NULL
NULL
NULL
NULL
AV NUTIBARA NO 26-31

Ilustración 2 Columna dirección de la tabla NIT, DB Ofimatica SA

Tabla 3

Propiedades de la tabla MTGLOBAL

Nombre	MTGLOBAL
Tamaño	994 Registros
Columnas	47
Tipo	Tabla de hechos
Uso	Transaccional
Versión DB	SQL Server

Tabla 3 MTGLOBAL DataBase Ofimatica SA

Otra de las tablas en la que se realizara el cambio es MTGLOBAL en la columna ORDEN o CODEMPRESA ya elegida las tablas a intervenir se procederá a realizar la transformación de la información con las herramientas Open Source.

Dominio de los datos.

El dato representa el conjunto de valores que puede obtener cada campo. El análisis genera reglas de datos conexas de forma directa con las reglas de negocio. Es útil establecer el dominio de los datos para establecer niveles de calidad según el cumplimiento de estas reglas y que generan perfilación y su correlación directa con los requerimientos de cada empresa.

Open Refine.

Creamos un nuevo proyecto en open refine. Los archivos a importar se pueden tener en formatos TSV, CSV, *SV, Excel (.xls y .xlsx), JSON, XML, RDF como XML. Datos en Documentos de Google también son compatibles. Compatibilidad para otros formatos puede ser añadida con las extensiones de Open Refine.

Elegimos el archivo TabalNIT.csv, para la correcta carga de la información, la codificación de los caracteres debe ser en UTF-8 para que los caracteres especiales las reconozcan y procedemos a crear nuestro proyecto.

Antes de todo para poder realizar las transformaciones de los datos debemos conocer que el lenguaje GREL, es propio de la herramienta para expresiones regulares, funciones, controles, funciones booleanas entre otros.

De la tabla NIT tomamos como muestra la columna dirección ya que esta muestra unos valores NULL, que pueden ser transformados y enriquecer los datos que se tiene, como se muestra en la siguiente imagen gráfica.

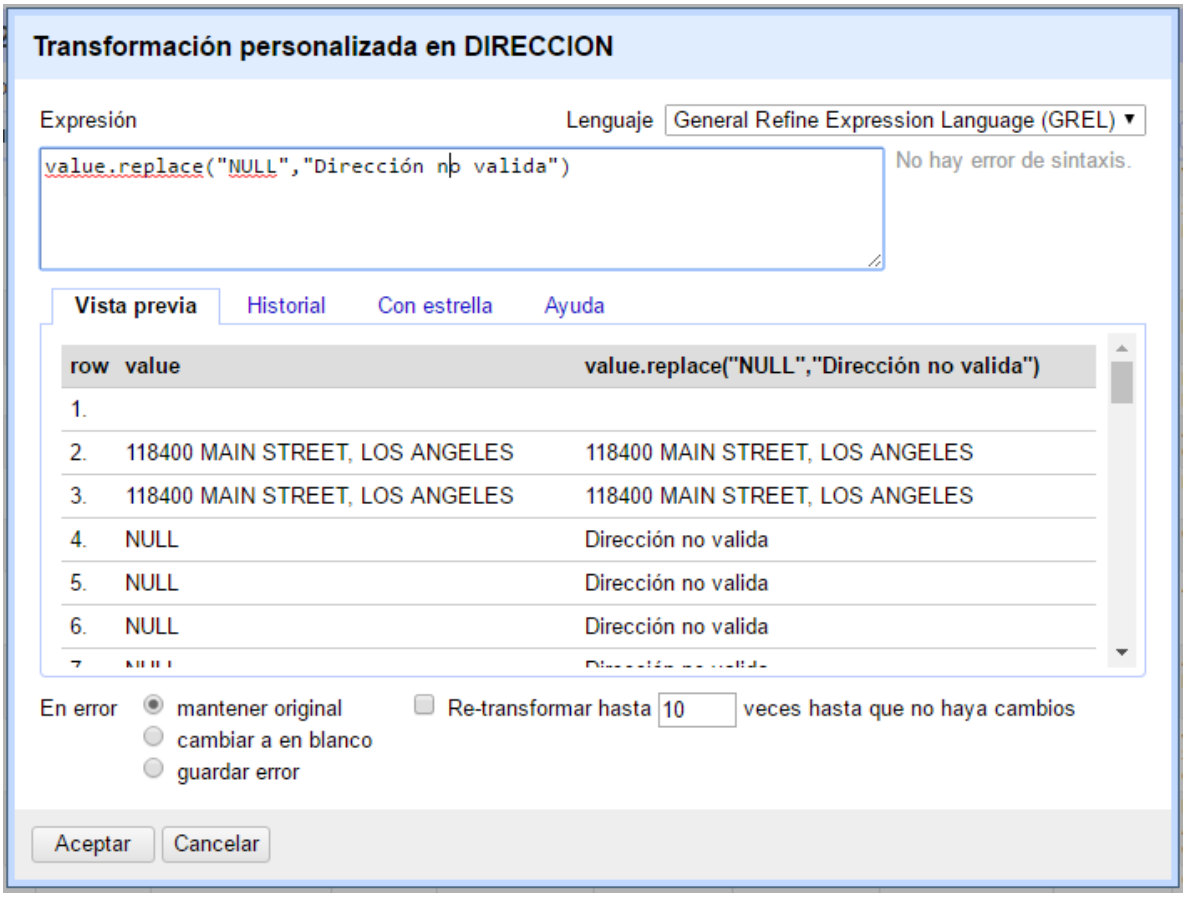


Ilustración 3 Transformacion personalizada columna dirección, tomada de aplicación Open Refine

En la ilustración 1, la expresión regular que se ingresó podemos convertir los datos en la columna que queremos reemplazar.

Vamos a crear otra expresión regular la columna NIT ya que esta no tiene un orden o formato.

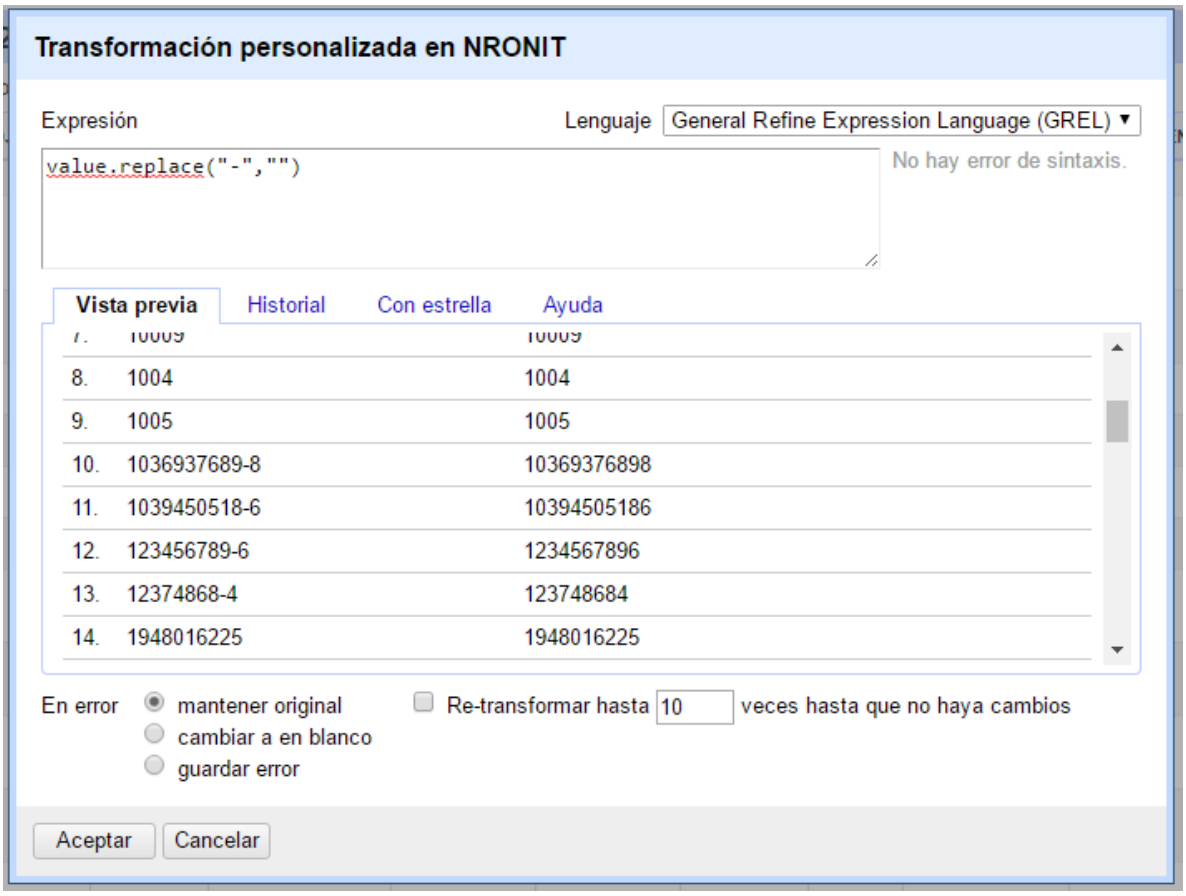


Ilustración 4 Transformacion perzonalida columna NRONIT, tomada de Open Refine.

En la ilustración 2, esta columna está correctamente organizada y con un formato adecuado, para que la calidad de los datos sea más adecuada para el cliente.

En la tabla MTBGLOBAL está la columna CODEMPRESA la cual aplicaremos una expresión regular value.replace("0","TODOTERRENO") y la transformación queda así:

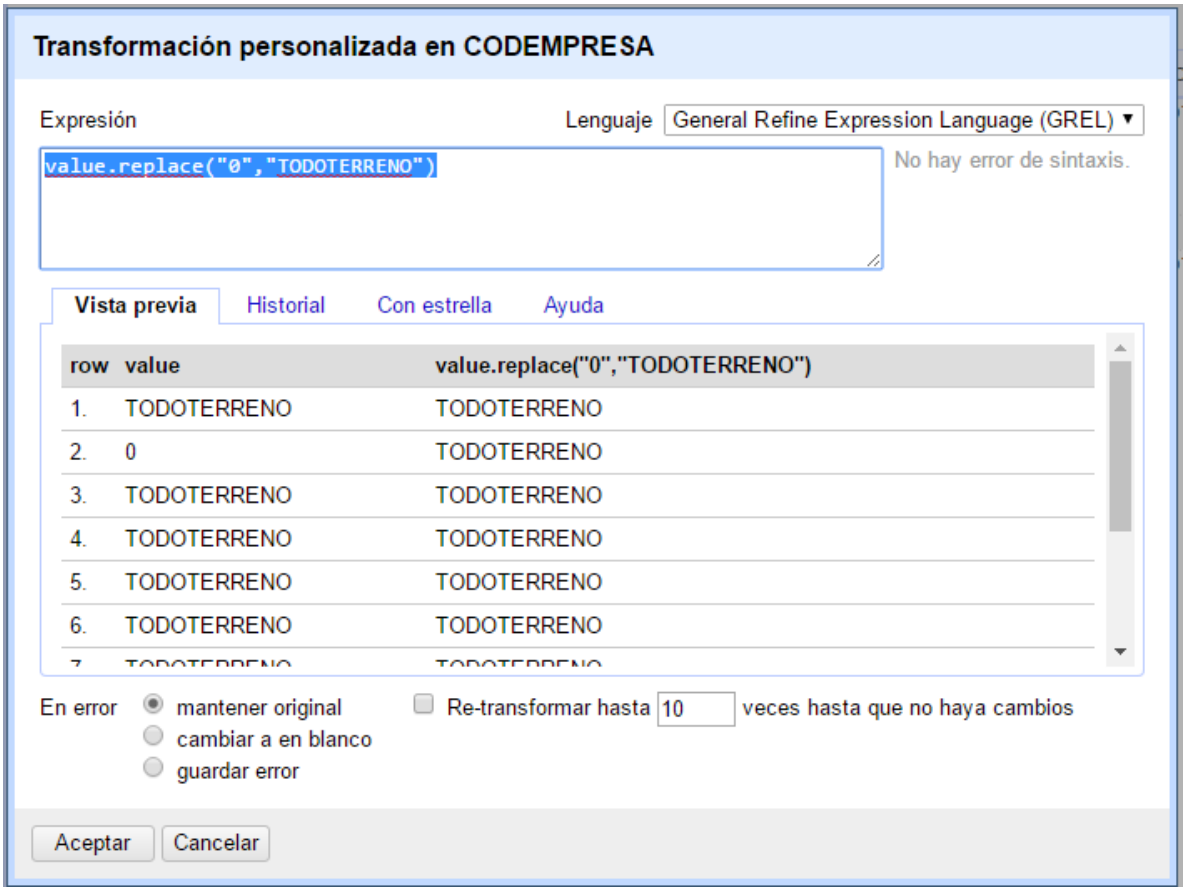


Ilustración 5 Transformacion personalizada columna CODEMPRESA, tomada de open refine.

En la ilustración 3, esta información tiene valores atípicos y con la expresión regular corregimos datos que son incorrectos como los vimos en la imagen anterior.

Open Talen Profiler

Se procede a realizar las pruebas para las tablas seleccionadas con la herramienta Talend open profiler, en este caso se inicia la experimentación con la tabla NIT específicamente para los valores de la columna DIRECCION.

Análisis principal Prueba 1:

Se crea dentro del nuevo proyecto un nuevo análisis para el cual es necesario seleccionar la tabla y columna de enfoque, lo que permitirá mostrar una opción con los diferentes indicadores de expresiones regulares y filtros creados previamente, que pueden ser aplicados en las columnas del análisis seleccionadas.

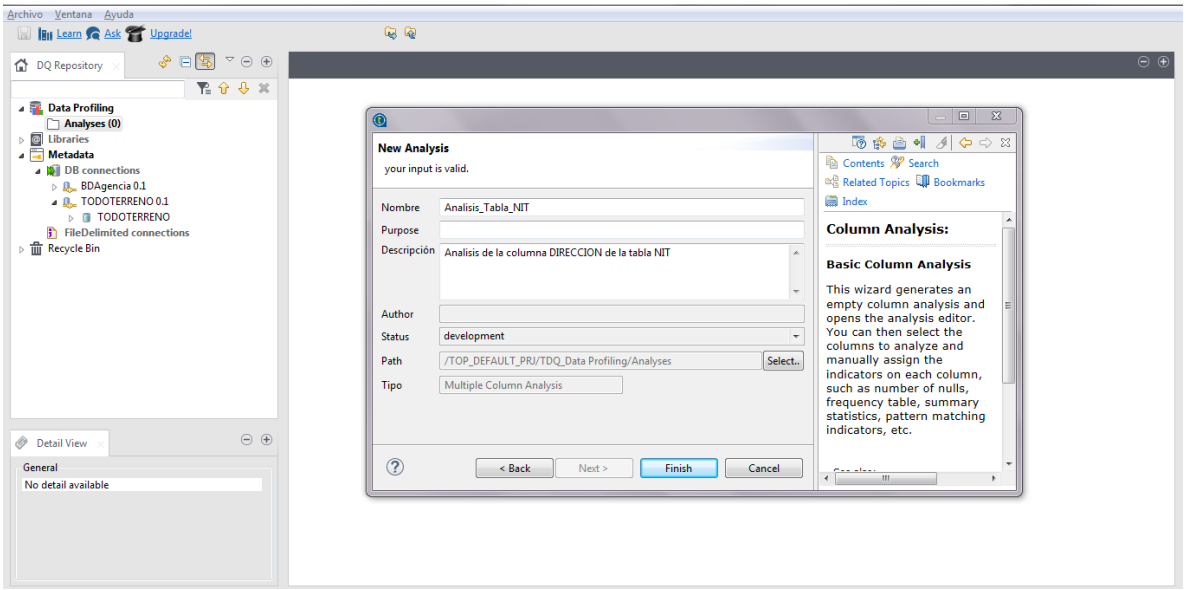


Ilustración 6 Creacion de un nuevo analisis, tomado de Talend Profile

Tal y como se presenta en la Ilustración 5, como siguiente paso se encuentra seleccionar la tabla y cada una de las columnas a las que se espera aplicar el análisis.

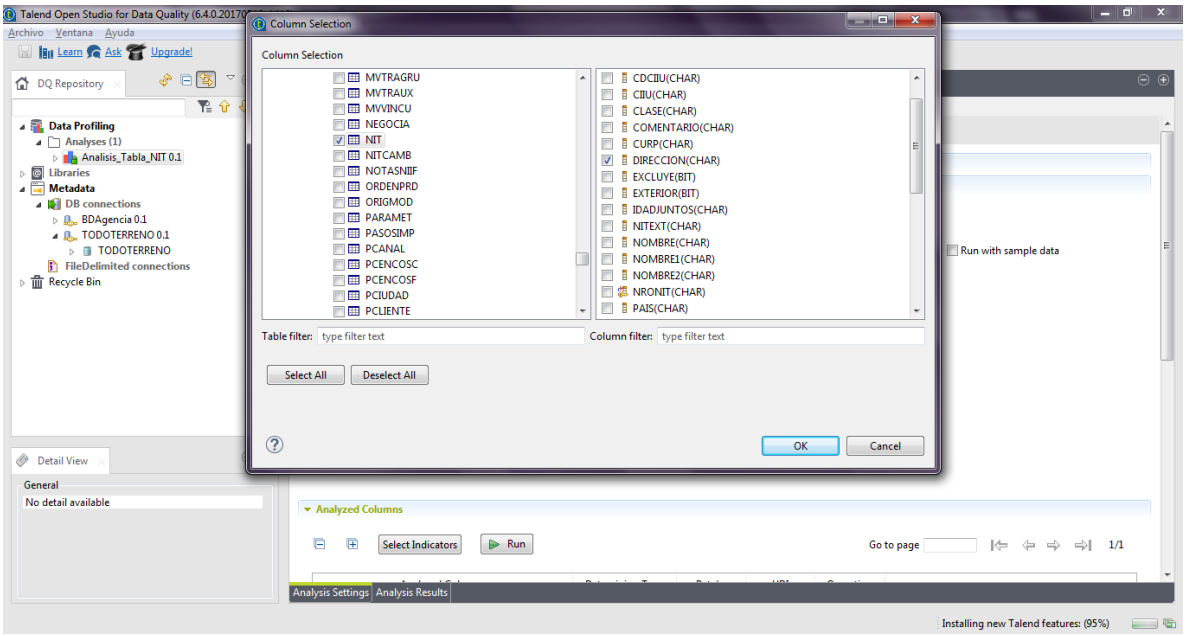


Ilustración 7 Selección de la tabla NIT, tomado de Talend Profile

Los datos de las columnas seleccionadas se listarán en el detalle del análisis y estará disponible una serie de opciones para aplicar sobre la columna en general, como el límite de datos a mostrar o los indicadores a aplicar, así como se muestra en la Ilustración 6.

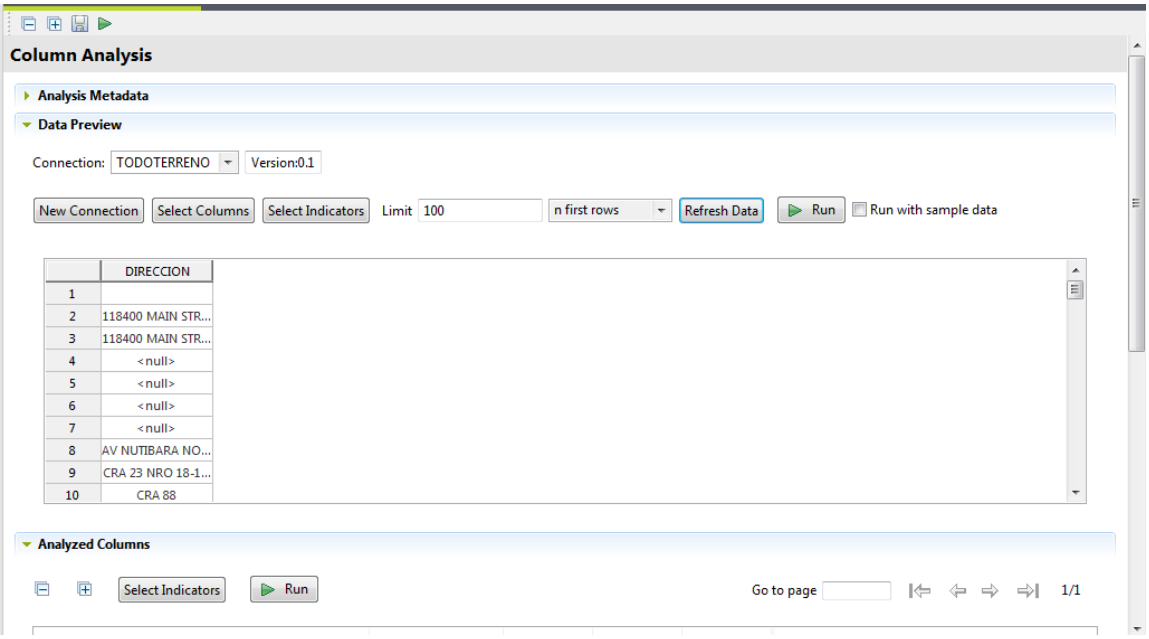


Ilustración 8 Analisis a la columna Direccion, tomado de Talend Profile.

Al seleccionar la opción de los indicadores se abre un nuevo detalle donde se especifica cada una de las expresiones regulares disponibles. Cada una de estas reglas se discrimina por categoría y afinidad, lo cual permite que la selección sea de manera más simple y puntual, tal y cómo puede visualizarse en la ilustración 7.

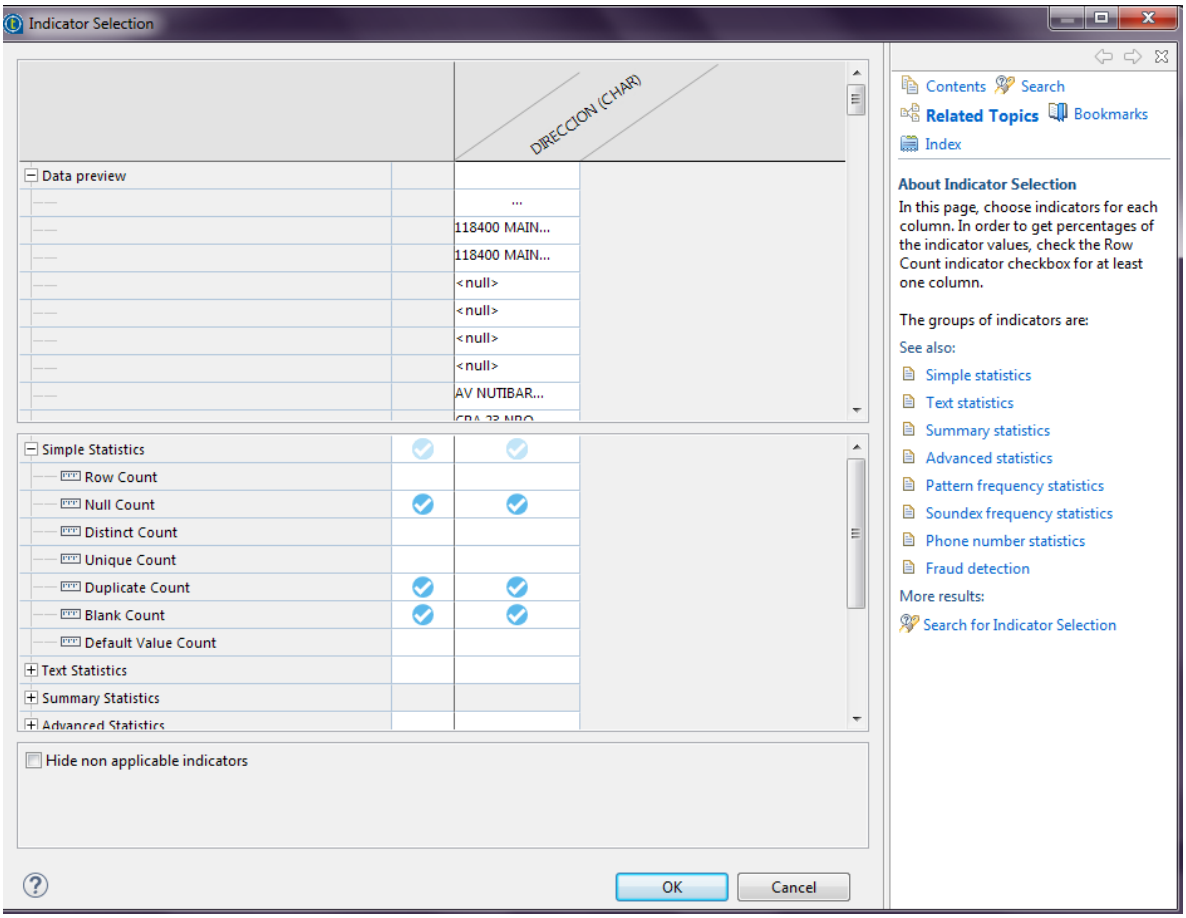


Ilustración 9 Reglas con la que se van a filtrar, Tomado de Talend Profile

Según las reglas o expresiones regulares seleccionadas, el objetivo principal es encontrar aquellos registros que estén vacíos, duplicados o en blanco:

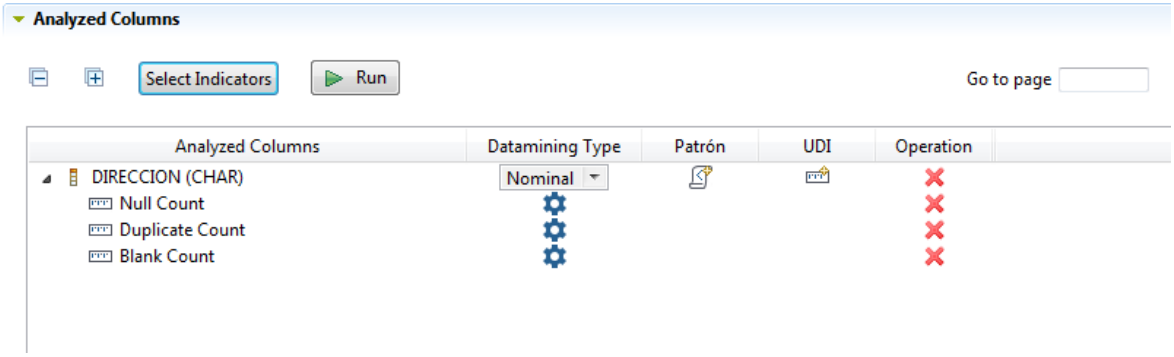


Ilustración 10 Reglas para aplicar expreciones regulares, tomado de Talen Profile

Después de terminar de configurar el análisis, se ejecuta y como resultado se obtienen los siguientes datos representados por el conteo de los datos.

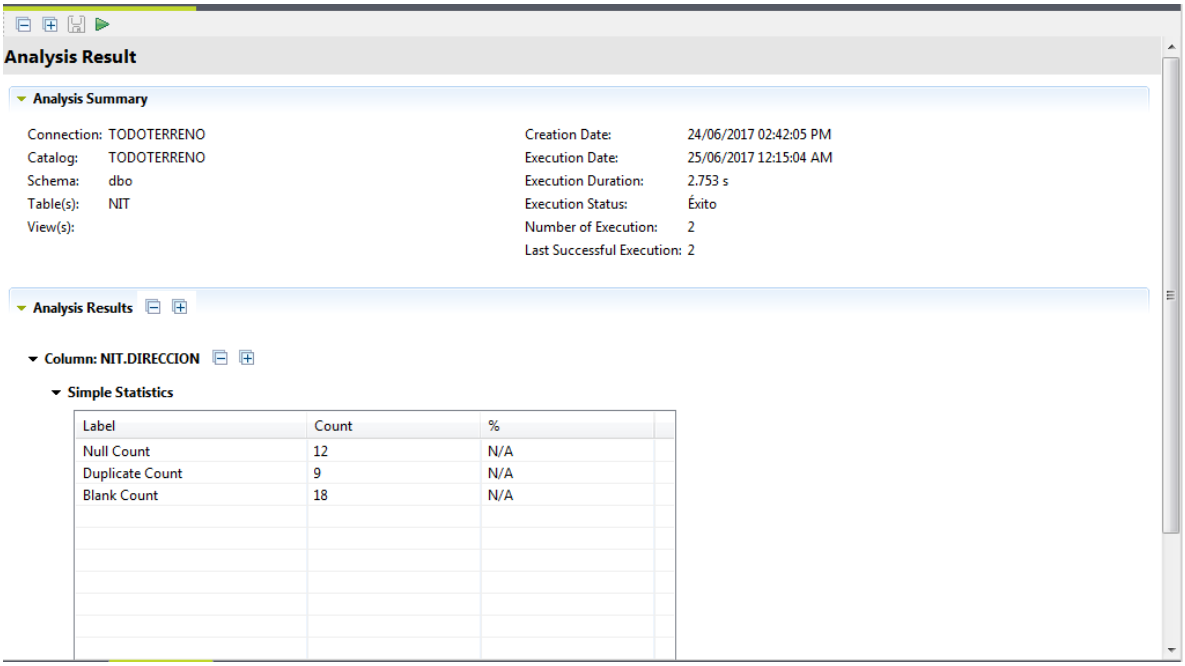


Ilustración 11 Resultados de la tabla NIT, tomado de Talend Profile.

Análisis adicional Prueba 1:

Este análisis se realiza con el recomendado por la misma herramienta, la cual presenta las posibles reglas que se pueden aplicar sobre el tipo de datos que posea la columna a analizar.

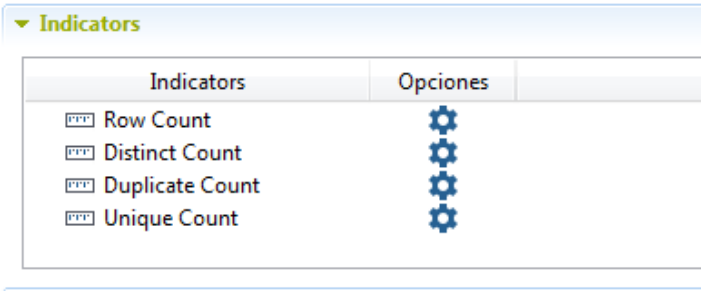


Ilustración 12 Reglas para aplicar expreciones regulares, tomado de Talen Profile

Los resultados son presentados de la siguiente manera, donde se discrimina una serie de estadísticas por la regla aplicada a los registros de la columna, mostrando un conteo de registros a los cuales aplica y el equivalente en porcentaje con respecto al total global de los registros detectados.

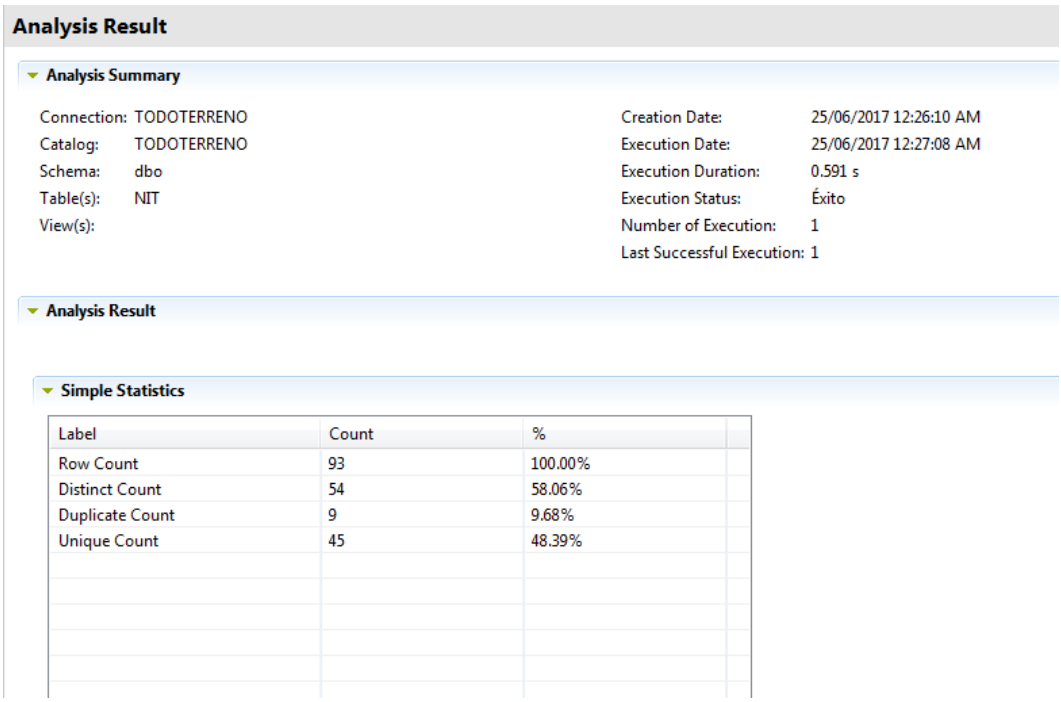
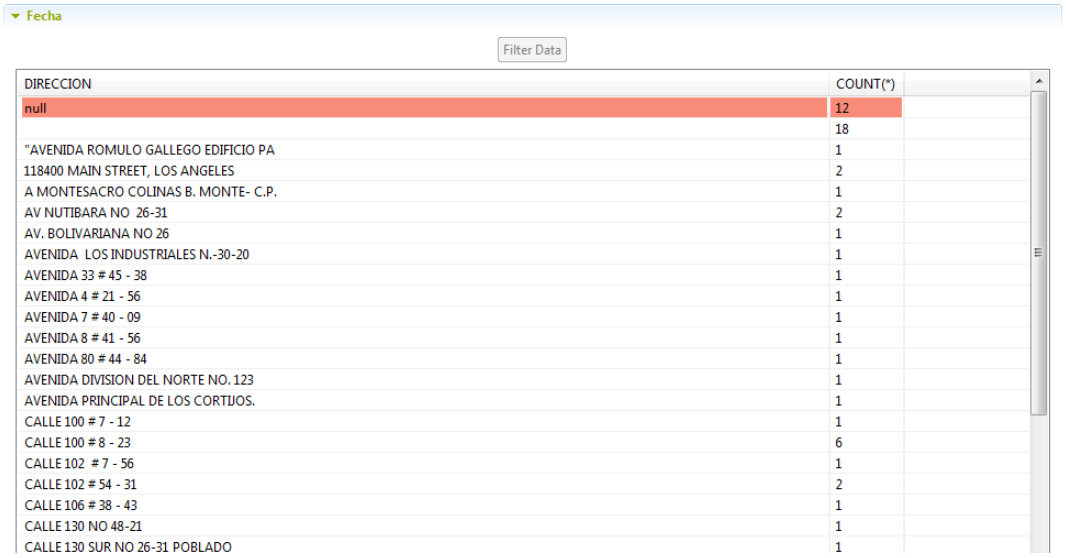


Ilustración 13 Resultados de la tabla Nit, Tomado de Talend Profile

Adicional a los resultados generales se presentan unos detalles a nivel de registros, donde se presenta un conteo por repetición de valores:



DIRECCION	COUNT(*)
null	12
	18
"AVENIDA ROMULO GALLEGO EDIFICIO PA	1
118400 MAIN STREET, LOS ANGELES	2
A MONTESACRO COLINAS B. MONTE- C.P.	1
AV NUTIBARA NO 26-31	2
AV. BOLIVARIANA NO 26	1
AVENIDA LOS INDUSTRIALES N.-30-20	1
AVENIDA 33 # 45 - 38	1
AVENIDA 4 # 21 - 56	1
AVENIDA 7 # 40 - 09	1
AVENIDA 8 # 41 - 56	1
AVENIDA 80 # 44 - 84	1
AVENIDA DIVISION DEL NORTE NO. 123	1
AVENIDA PRINCIPAL DE LOS CORTIJOS.	1
CALLE 100 # 7 - 12	1
CALLE 100 # 8 - 23	6
CALLE 102 # 7 - 56	1
CALLE 102 # 54 - 31	2
CALLE 106 # 38 - 43	1
CALLE 130 NO 48-21	1
CALLE 130 SUR NO 26-31 POBLADO	1

Ilustración 14 Resultados Generales a nivel de registros, tomado de Talend Profile.

Análisis Tabla Mtglobal

Como segunda parte de la experimentación con la herramienta, se procede a cambiar el análisis con la columna Orden de la tabla MTGLOBAL, la cual posee diferente tipo de datos y que a su vez se puede aplicar las mismas reglas de calidad de datos.

Análisis principal Prueba 2:

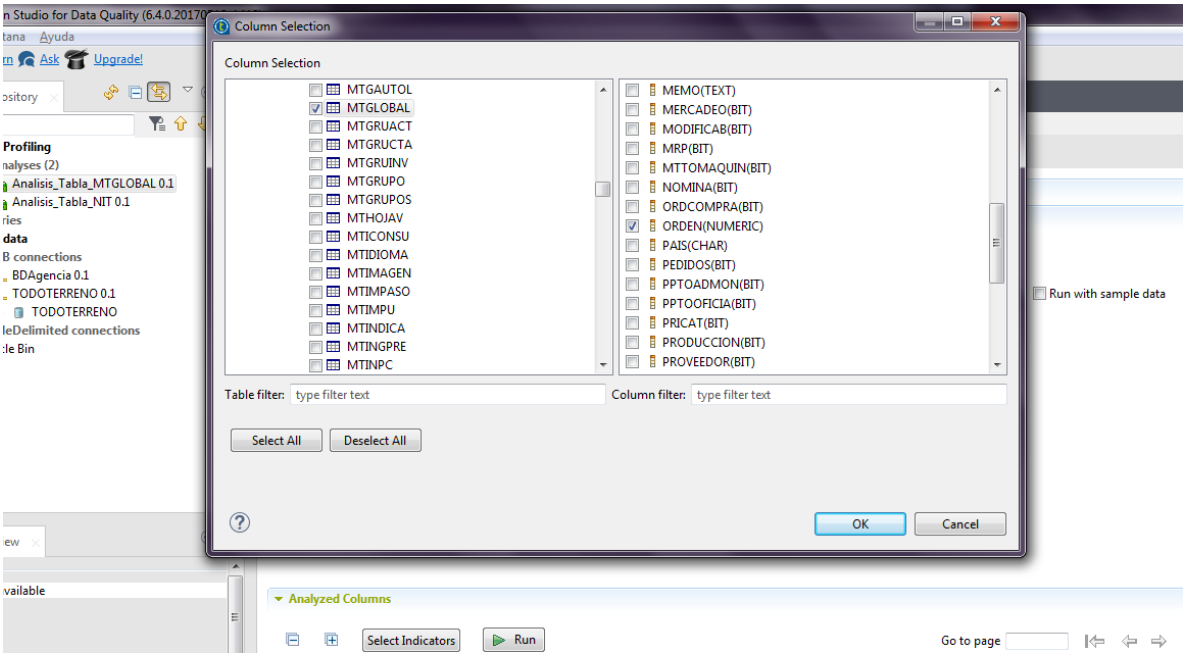


Ilustración 15 Selección de tabla MTLGLOBAL

La lista de datos es presentada, para la cual está disponible diferentes opciones de configuración de cantidad de datos e indicadores a aplicar como parte de la misma prueba.

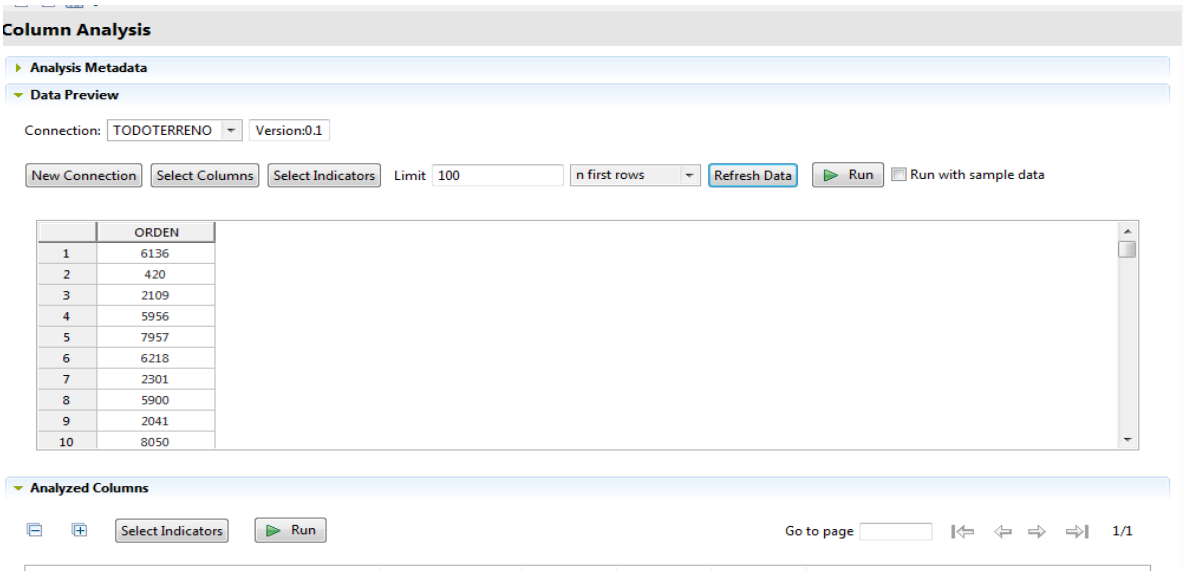


Ilustración 16 selecciona tabla Orden, tomado de Talend Profile.

Una vez se tiene la cantidad de datos a utilizar, se ingresa al detalle de los indicadores o expresiones regulares, como se muestra en la siguiente imagen:

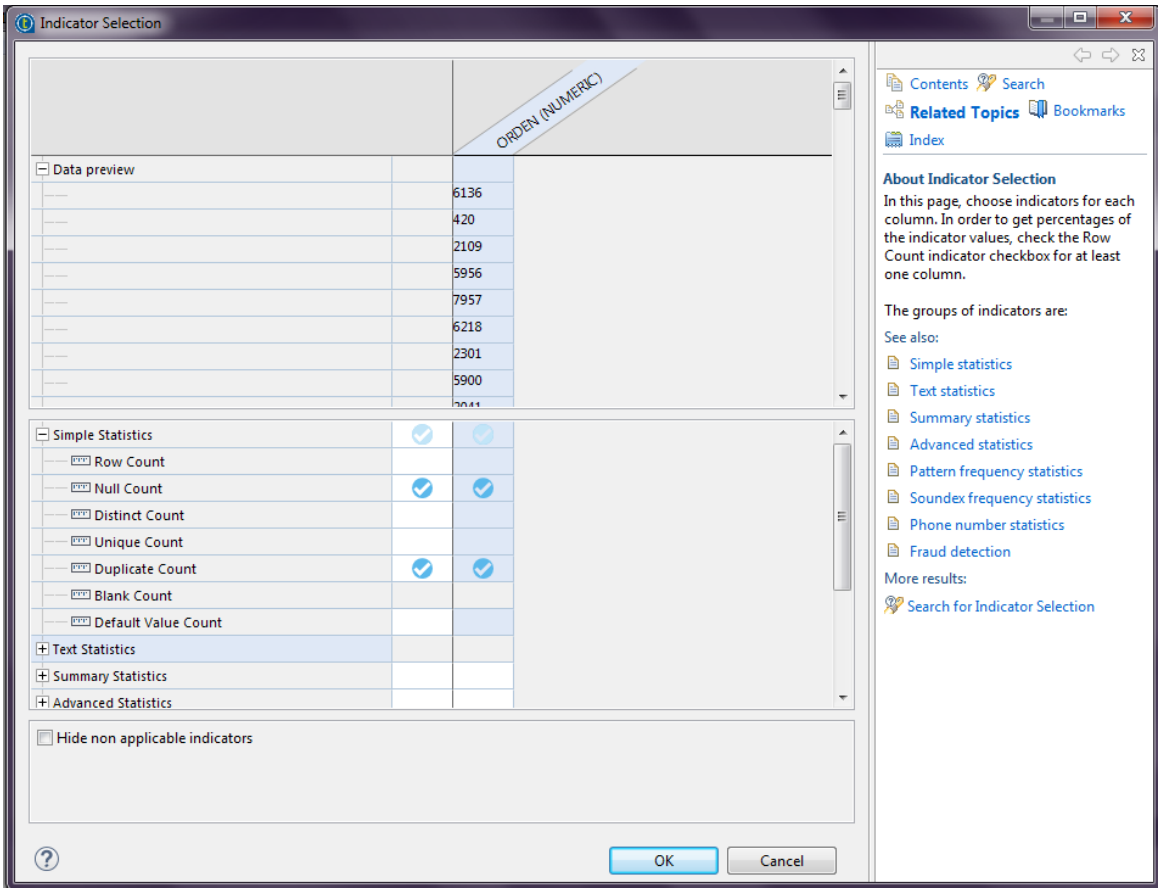


Ilustración 17 Indicadores de expreciones regulares, tomado de Taled Profile.

Tras el proceso realizado como objetivo principal de la prueba es encontrar aquellos registros que estén vacíos o duplicados:

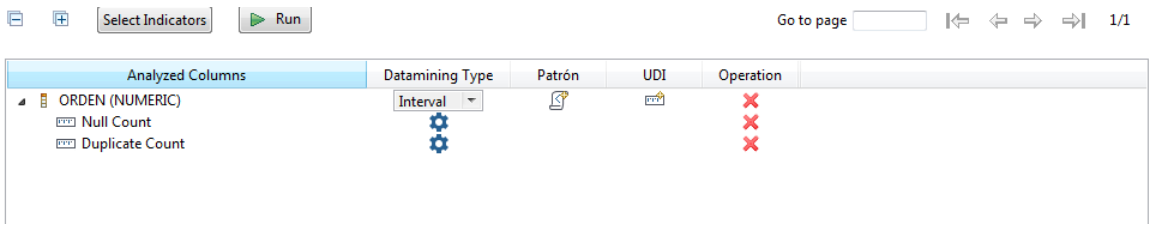


Ilustración 18 Aplicar expreciones regulares, tomado de Taled Profile

Luego de realizar la configuración correspondiente del análisis se reciben los siguientes resultados, donde principalmente se representa en un conteo el total de los registros analizados a los que le aplica las reglas establecidas.

Resultados:

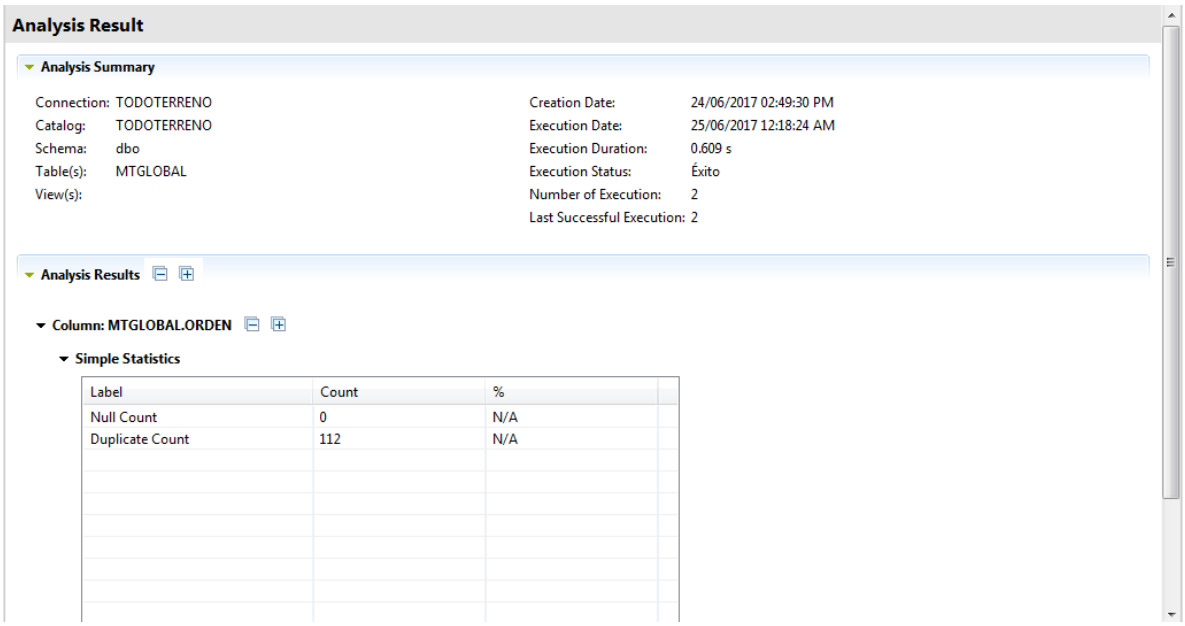


Ilustración 19 Resultados del analisis, tomado de Talend Profile

Análisis adicional Prueba 2:

Al igual que en la anterior prueba 1 es importante realizar el análisis con las expresiones regulares que son sugeridas por la herramienta en cuanto a calidad del tipo de dato se refiere.

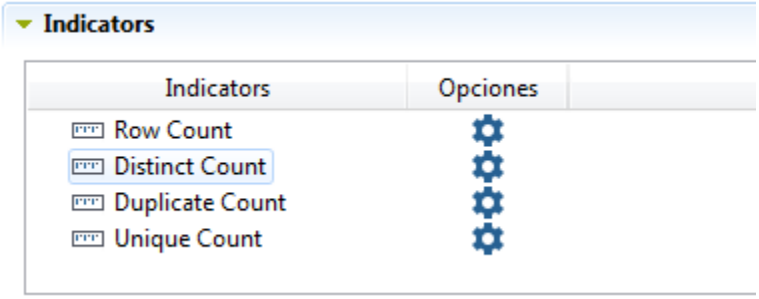


Ilustración 20 Analisis de las expreciones regulares, tomado de Talend Profile

Una vez se termina de configurar el análisis y se verifican todas las posibles opciones a emplear, es pertinente ejecutar el proceso de calidad de datos. Los resultados que se obtienen durante dicho proceso se ven reflejados en el siguiente conteo y su correspondiente representación en porcentajes sobre el total de registros encontrados luego de aplicar las reglas de calidad de datos.

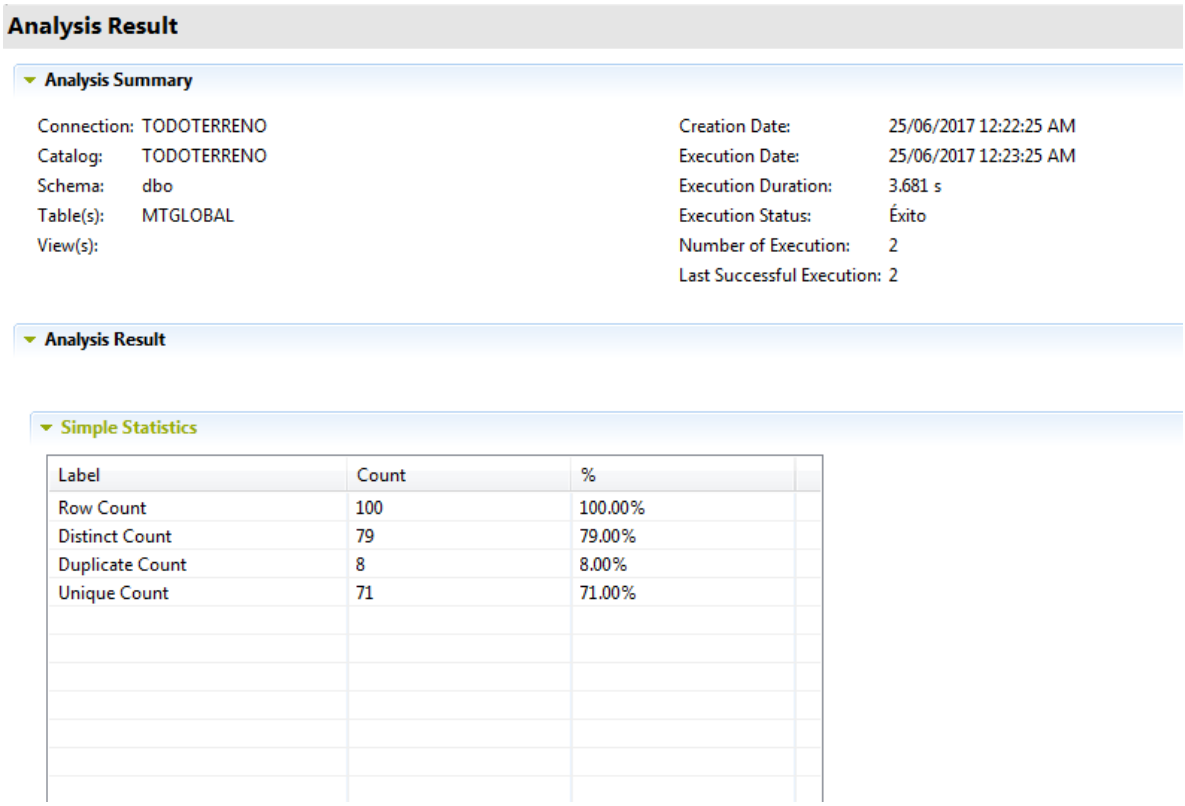


Ilustración 21 Resultados finales, tomado de Talend Profile.

SQL Porwer DQGuru.

Para poder utilizar esta herramienta fue necesario convertir el BD de SQL a MySQL, con esta migración podemos utilizar esta, creamos el proyecto “Análisis”, configuramos la base de datos con una conexión JDBC.

Para poder trabajar invocamos la tabla NIT y creamos la expresión regular que mostraremos a continuación.

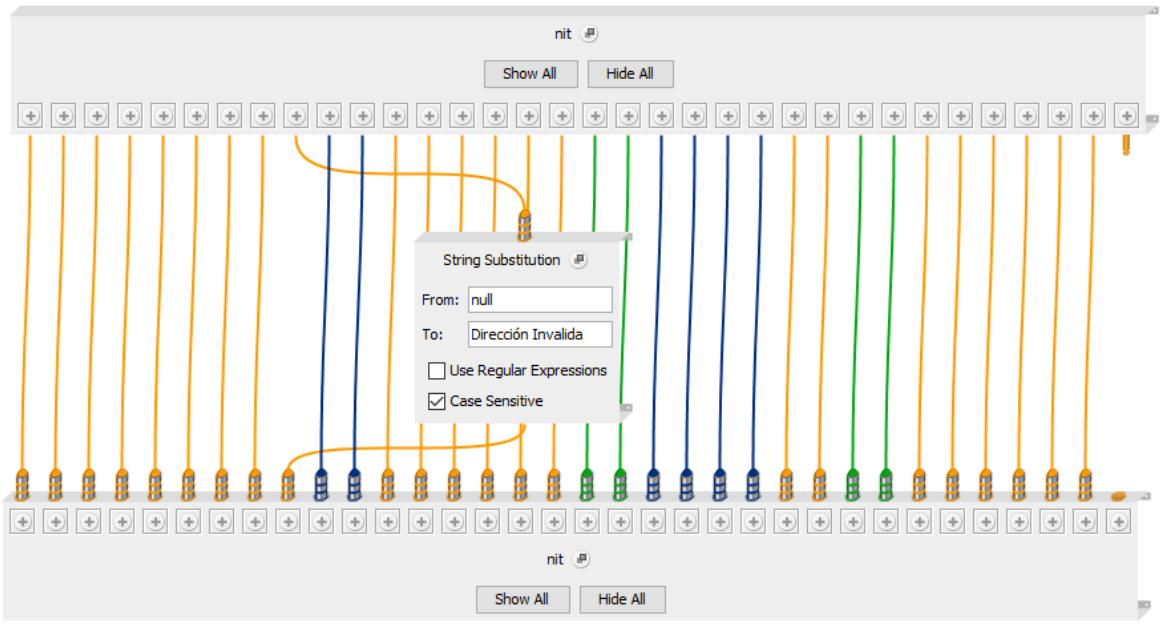


Ilustración 22 Transoformacion de valores, tomada de SQL DQGuru.

En la ilustración 20, la transformación pasamos de valores NULL al valor “Dirección invalida”.

Creamos otra regla con expresiones regulares para transformar la tabla MTGLOBAL.

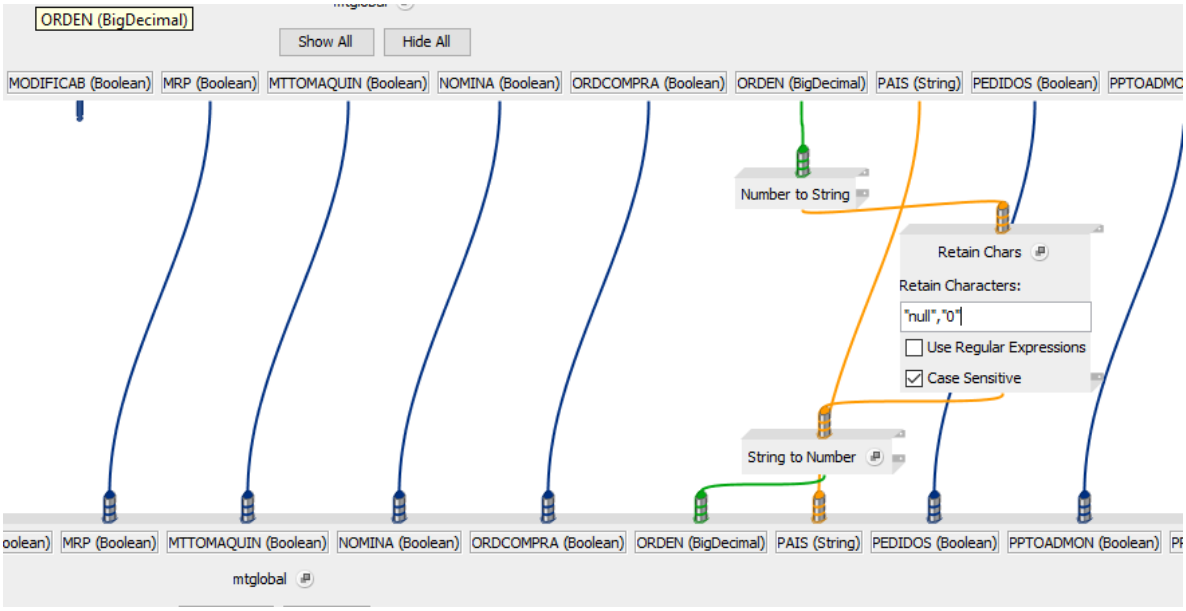


Ilustración 23 Cambio de valores, tomada de SQL DQGuru

En esta ilustración 21, esta regla fue necesario tener en cuenta que los valores por defecto venían numéricos, se realizó una conversión interna para que reconozca los valores null de la tabla y a su vez se convirtió de texto a número para el dato no tenga ningún problema al transformar.

3. RESULTADOS Y DISCUSIÓN

El proceso de perfilación de datos permite conocer a fondo los datos de estudio con base en las estadísticas que generó Open Talend Profiler, luego de la tabulación y el ordenamiento automatizado por expresiones regulares.

“Examina si los datos de una organización son confiables, consistentes, actualizados, están libres de duplicidades y si son apropiados para sus objetivos” (Vilalta Alfonso, 2018). A partir de los hallazgos encontrados, se tiene que las cualidades de cada una de las herramientas son diferentes en el modo de actuar, pero en su filosofía tiene el mismo fin, como lo es la calidad de los datos que resulta de todo un proceso de análisis y ejecución de cada una de estas.
Según la naturaleza de las reglas de negocio obtenidas de los requerimientos, se define que la dimensión de calidad asociada a dichas reglas es la exhaustividad. Que indica que mientras más información se obtenga en la captura de datos, mayor será la calidad de dichos datos.

Pero en lo que no estamos de acuerdo es que en las empresas que tienen datos optan por una herramienta de pago y eso es lo que hemos encontrado en su mayoría de autores, por eso hacemos énfasis en este trabajo de grado que hay aplicaciones open source con muchas más características y de fácil acceso para que en la toma de decisiones sea correcta para la empresa que desee utilizar estas.

Para demostrar de qué características tiene cada una de estas realizamos una medición cualitativa que lo mostramos en la siguiente tabla:

TABLS 4

MATRIZ CUALITATIVA

Herramientas de Calidad de Datos	**Matriz Cualitativo**				
	Integración con Motores de BD	Perfilamiento de datos	Genera archivos de salida, CVS y otros	Multiplataforma	Transformaciones de datos.
Power DQGuro	X		X	X	X
Talend Open Profiler	X	X	X	X	X
Open Refine			X	X	X

Tabla 4 Matriz Cualitativa Construcción propia

Tratamiento y transformación de datos son las bases principales de cada una estas herramientas, que a través de una experimentación se ve como resulto una matriz cualitativa donde se plantean cada una de las características que más predominan en cada una de ellas. Como una de las principales cualidades se destaca la versatilidad y eficiencia en realizar su integración con uno o varios motores de DB, lo cual es importante para las PYMES ya que las fuentes de datos no se alterarían tan fácilmente por el manejo independiente de su origen, posteriormente arrojando resultados basados en las reglas o decisiones aplicadas sobre la herramienta y la potencial explotación del resto de beneficios proporcionados y que se acoplen mejor a sus necesidades de negocio.

4. CONCLUSIONES, RECOMENDACIONES Y TRABAJO FUTURO

Conclusiones

- El conjunto de datos elegido como fuente de la investigación metodológica fue acorde a los objetivos planteados, ya que permitió el correcto desarrollo de la búsqueda y presentación de la metodología.
- El perfilado de los datos permitió la obtención de un conocimiento más adecuado, acerca del conjunto de datos a estudiar. Por medio de este proceso, fue posible obtener las reglas de datos.
- Es posible concluir que las empresas pueden apoyarse en este tipo de plataformas open source que proporcionan en gran medida los principales elementos para llegar a ser más eficaces los análisis de datos antes de realizar un proceso de inteligencia de negocios, ya que, a mayor filtro, mayor la limpieza de la información y, por ende, mejor la calidad obtenida para una adecuada toma de decisiones.
- En la evaluación realizada a nivel de cada herramienta, se destaca el cumplimiento de los requisitos mínimos que cada empresa puede tener en cuanto a un adecuado análisis en calidad de datos se refiere.
- La dimensión de exactitud, por su naturaleza, es muy difícil de medir ya que se deberá comprobar campo por campo si su información es verdadera, de acuerdo con el referente.
- De cada una de las herramientas evaluadas, tanto las características como las cualidades que poseen, permite que las empresas puedan tener un fuerte apoyo en tema de calidad de datos dependiendo principalmente de las necesidades puntuales existentes, para poder llevar mejor un análisis y así tomar mejores decisiones.

Recomendaciones.

- Los estudiantes en carreras afines a sistemas y futuros administradores de bases de datos o que estén en inteligencia de negocios, deben profundizar en los temas de data governance, data quality y big data. Por alguna razón válida, en los países líderes en tecnología de punta, toman muy en cuenta estos aspectos para tenerlos en consideración a la hora de implementar o dar mantenimiento a sistemas, no solo analíticos, sino también transaccionales.
- Los docentes encargados de las actualizaciones de los currículos de estudios de las diferentes carreras afines a sistemas, deben incluir activamente los temas del punto anterior, para cumplir con sus propios trabajos derivados de su quehacer profesional y para dotar a los estudiantes de las habilidades que surgen al utilizar las herramientas open source actualizadas y que están disponibles en el mercado, para formar profesionales competentes y con una actuación de capacidad que conlleve el éxito en el ejercicio profesional. Un reto importante es lograr una aceptación de la institución (ITM) para aprovechar la amplia gama de posibilidades y recursos que proporciona el usuario de estas herramientas open source.

Trabajos Futuros.

Finalmente, se propone profundizar en el estudio del ciclo de vida de calidad de datos desde un enfoque teórico para analizar mejor el proceso aplicado por las herramientas de perfilamiento de datos, posibilitando esclarecer conceptos desde el ciclo hasta la experimentación además de mantenerse actualizado en las herramientas ya que estas van evolucionando al pasar el tiempo.

REFERENCIAS

Bibliografía

S. Wilke, A. M. (2014). *A framework for assessing the quality of aviation safety databases.*

Schoenbach, V. J. (2004). Gestión y análisis de datos. *Comprendiendo los fundamentos de la epidemiología— un texto en desarrollo*.

Lopez, C. P. (2007). *Minería de datos, técnicas y herramientas.*

López, C. P. (2008). *Minería de datos: técnicas y herramientas.*

Sinnexus. (s.f.). *http://www.sinnexus.com/business_intelligence/datamining.aspx.*

Valles, J. (29 de 12 de 2013). *Marketing Digital*. Obtenido de https://josepvalles.wordpress.com/2013/12/29/chuleta-guia-tutorial-open-refine-google-refine/

J.Maynard-Smith. (1982). *Dictionary of data processing.*

Valdés, D. P. (26 de 10 de 2007). *Maestros de la Web*. Obtenido de Maestros de la Web: http://www.maestrosdelweb.com/que-son-las-bases-de-datos/

Rouse, M. (s.f.). *TechTarget*. Obtenido de 2015

PowerData. (26 de Abril de 2016). *Cómo puede ayudar el perfilado de datos al BI*. Obtenido de http://blog.powerdata.es/el-valor-de-la-gestion-de-datos/como-puede-ayudar-el-perfilado-de-datos-al-bi

Rios, M. (30 de Junio de 2014). *Escuela de datos*. Obtenido de https://es.schoolofdata.org/2014/06/30/openrefine/

SQLPower. (2017). *Data Cleansing & Address Correction: SQL Power DQguru*. Obtenido de SQLPower: http://www.sqlpower.ca/page/sysreq

Talend. (2017). *Talend*. Obtenido de specifications data quality: https://www.talend.com/products/specifications-data-quality

Gardey, J. P. (2009). *Definicion*. Obtenido de Definicion: https://definicion.de/pyme/

Vilalta Alonso, J. A. (2008). *Metodología para el diagnóstico de la calidad de los datos.*

Vilalta Alonso, J. A. (2008). Metodología para el diagnóstico de la calidad de los datos.

Vilalta Alfonso, J. A. (2018). Metodología para el diagnóstico de la calidad de los datos.

APÉNDICE

Institución Universitaria	INFORME FINAL DE TRABAJO DE GRADO	Código	FDE 089
		Versión	03
		Fecha	2015-01-22

FIRMA ESTUDIANTES Natalia Andrea Jaramillo.
Carlos Eduardo Ossa Quintero
Mile Yurley Orrego Porras

FIRMA ASESOR _____

FECHA ENTREGA: 6/03/2018.

FIRMA COMITÉ TRABAJO DE GRADO DE LA FACULTAD _____

RECHAZADO___ ACEPTADO___ ACEPTADO CON MODIFICACIONES_____

ACTA NO._____

FECHA ENTREGA: _____

FIRMA CONSEJO DE FACULTAD_____

ACTA NO._____

FECHA ENTREGA: _____

47