Documentación Herramienta DataCleaner

Fecha	25/07/2019
Autor	Javier Belmar Tevar
Objetivo	Elaborar la documentación de la herramienta DataCleaner

ÍNDICE

1 INTRODUCCIÓN

DataCleaner es una herramienta que sirve para analizar la calidad de los datos obtenidos, con capacidad para encontrar patrones y supervisar los valores de los datos. Está construida para poder manejar pequeñas y grandes cantidades de datos. Es posible diseñar nuestras propias reglas de limpieza de datos y componerlas en múltiples escenarios distintos o bases de datos objetivo, dichas reglas pueden ser: reglas de búsqueda y/o reemplazo, expresiones regulares, coincidencia de patrones (pattern matching) o transformaciones totalmente personalizadas.

Ofrece también un ecosistema de integraciones de extensiones de aplicaciones impulsadas por la comunidad, contenido compartido, etc. Es posible hacer que DataCleaner funcione Hadoop y Apache Spark, además es viable integrarlo en otras aplicaciones como Pentaho Data Integration.

2 DATA QUALITY, DATA PROFILING Y DATA WRANGLING

En DataCleaner estos 3 conceptos están muy presentes por lo que se hará una breve explicación de cada uno de ellos.

2.1 Data quality

Data quality es un término que se refiere muchas veces a la calidad de los datos utilizados en las decisiones de negocio. Ejemplos de problemas de calidad de datos pueden ser: Completitud de los datos, corrección de los datos, duplicidad de datos y estadarización de los datos. **Dentro de Data quality tenemos el análisis de la calidad de los datos (DQA)** que, desde un punto de vista técnico, la tarea principal en un DQA es la actividad de creación de perfiles de datos (**Data profiling**), que le ayudará a descubrir y medir el estado actual de las cosas en los datos.

2.2 Data profilingPasos a realizar en el Nodo Coordinador/Maestro

El Data profiling es la actividad de investigar un almacén de datos para crear un “perfil” de él, para así poder hacer un mejor uso y mejoras. Posteriormente se introducirá el Data monitoring para poder monitorizar (comprobar) los datos medidos en el data profiling. Dentro de DataCleaner, un analizador es un componente que inspecciona los datos que recibe y genera un resultado o un informe. La mayoría de las actividades de Data profiling se desempeñan con los **Analizadores**.

2.3 Data wrangling

Es un proceso de conversión o asignación manual de datos de un formulario "en bruto" a otro formato que permite un consumo más conveniente de los datos con la ayuda de herramientas semiautomáticas. Esto puede incluir más información, visualización de datos, agregación de datos, capacitación de un modelo estadístico, así como muchos otros usos potenciales. Estas actividades se realizan principalmente en los **Transformadores.**

Más adelante se mostrarán ejemplos de Analizadores y Transformadores.

3 INSTALACIÓN, USO Y CONCEPTOS DE DATACLEANER

3.1 Conceptos

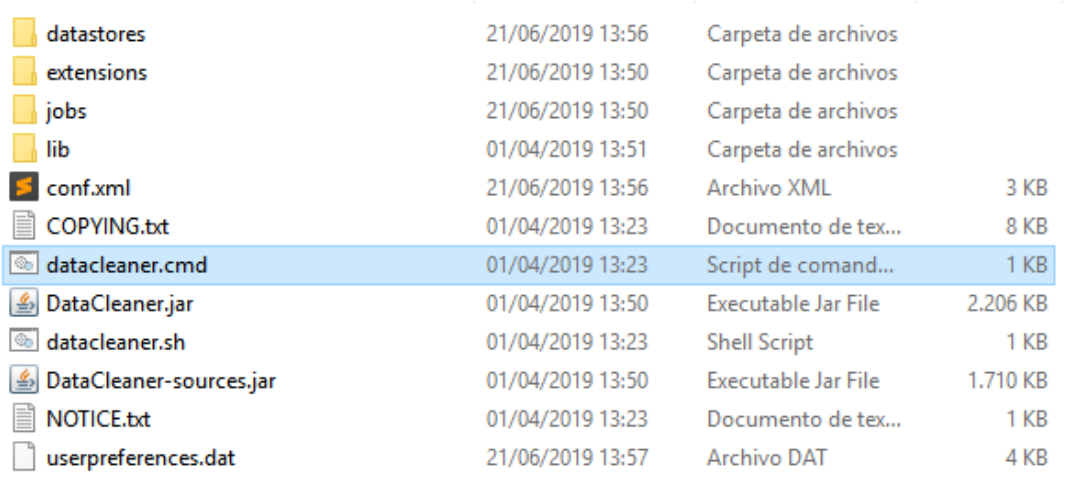
El término que usa DataCleaner para referirse a una fuente de datos es Datastore.

Almacén de datos compuesto (composite datastore): Un almacén de datos compuesto contiene múltiples almacenes de datos. La principal ventaja de un almacén de datos compuesto es que le permite analizar y procesar datos de múltiples fuentes en el mismo job.

3.2 Instalación

Prerrequisitos: Es necesario Tener un JRE con versión 7 o superior e interfaz gráfica.

Al descargar el zip y descomprimirlo obtendremos la carpeta que se muestra a continuación:



datacleaner.cmd, que contiene lo siguiente, nos permitirá ejecutar DataClenaer.jar:

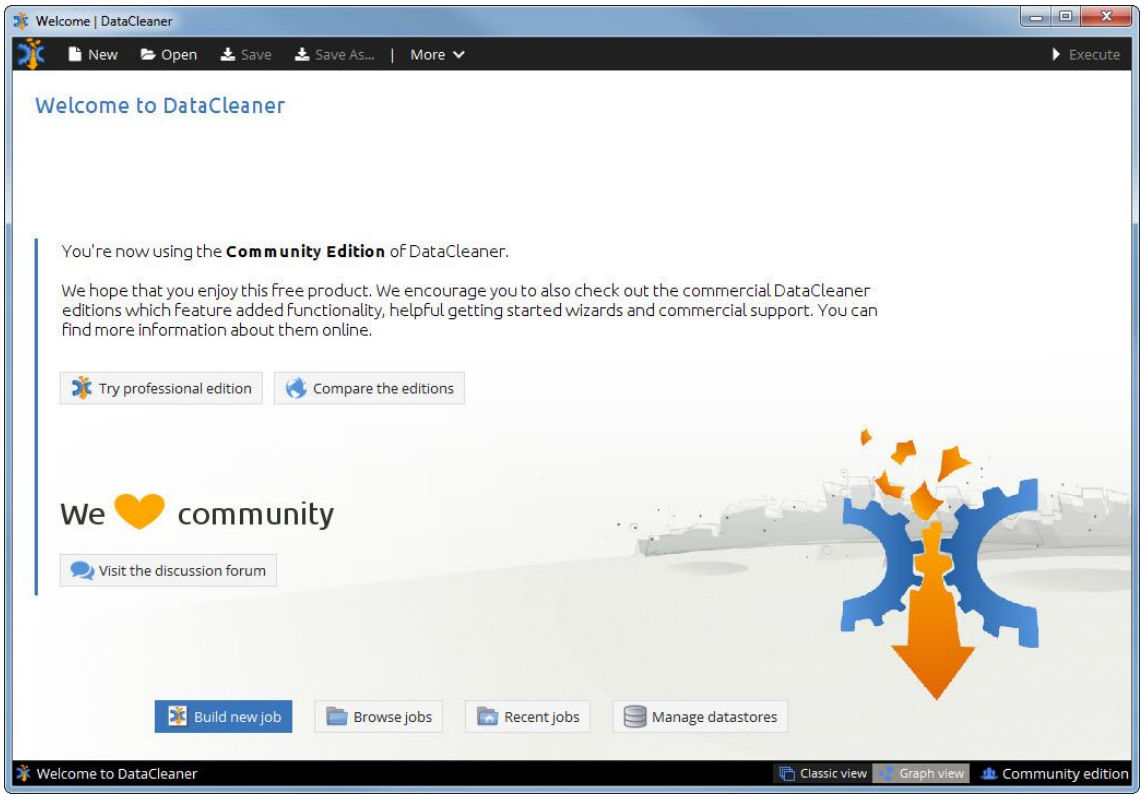
```
@echo off

set DATACLEANER_HOME=%~dp0

echo Using DATACLEANER_HOME: %DATACLEANER_HOME%

set DATACLEANER_LIB_HOME=%DATACLEANER_HOME%

echo Using DATACLEANER_LIB_HOME: %DATACLEANER_LIB_HOME%

cd /d %DATACLEANER_HOME%

set DATACLEANER_JAVA_OPTS=%JAVA_OPTS% -Xmx1024m

echo Using DATACLEANER_JAVA_OPTS=%DATACLEANER_JAVA_OPTS%

call java %DATACLEANER_JAVA_OPTS% -jar %DATACLEANER_LIB_HOME%\DataCleaner.jar"
```

Como vemos, hace uso de variables para lib y establece un home principal.

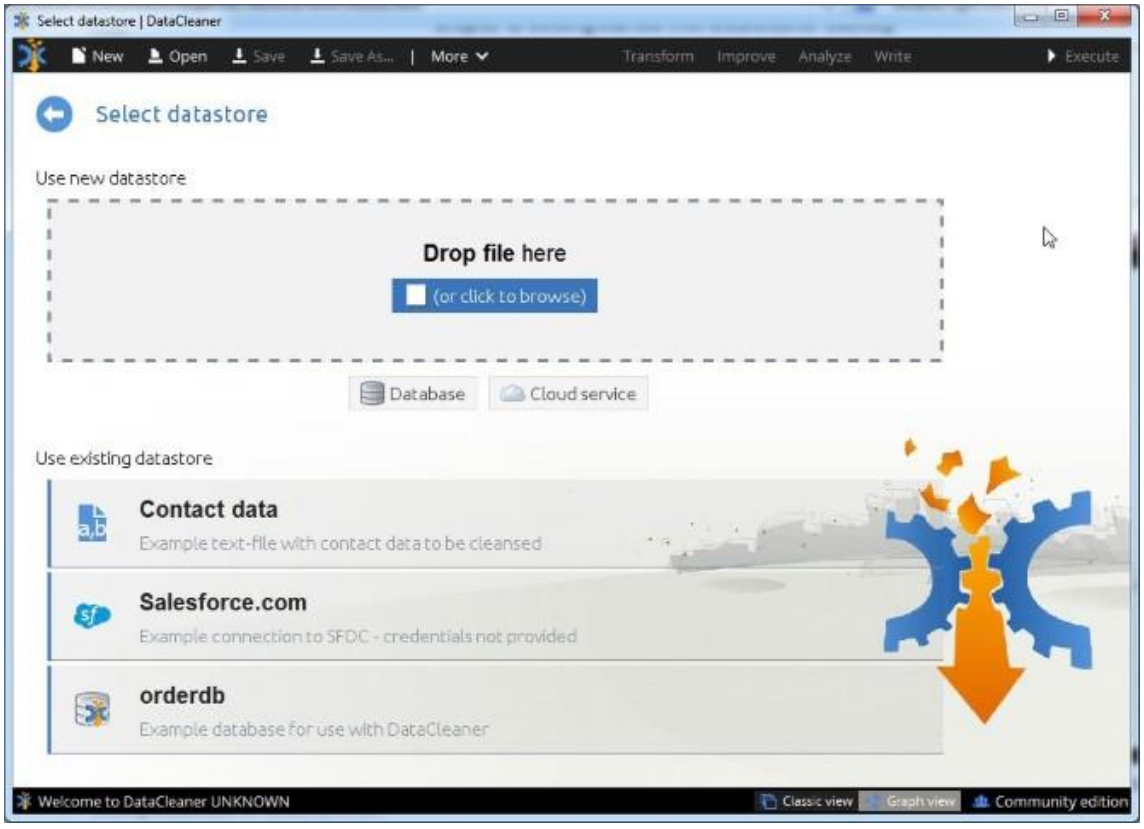
Para iniciar DataCleaner ejecutamos el .cmd (también hay un .sh).

3.3 Inicio

La imagen de abajo muestra la pantalla que aparece al lanzar DataCleaner. Podemos añadir nuevos datastores (fuentes de datos) pulsando sobre build new job.



Al pulsar sobre build new job nos aparecerá la siguiente pantalla.



Podemos arrastrar ficheros como CSVs o añadir directamente una base de datos o un cloud service. Los formatos de los archivos arrastrados en la drop zone serán inferidos, si se necesita algo más específico y personalizable como fuente de datos, en la pantalla de bienvenida está la opción de Manage datastores que nos permite más control de sobre como el fichero puede ser interpretado.

3.4 Componentes de un job

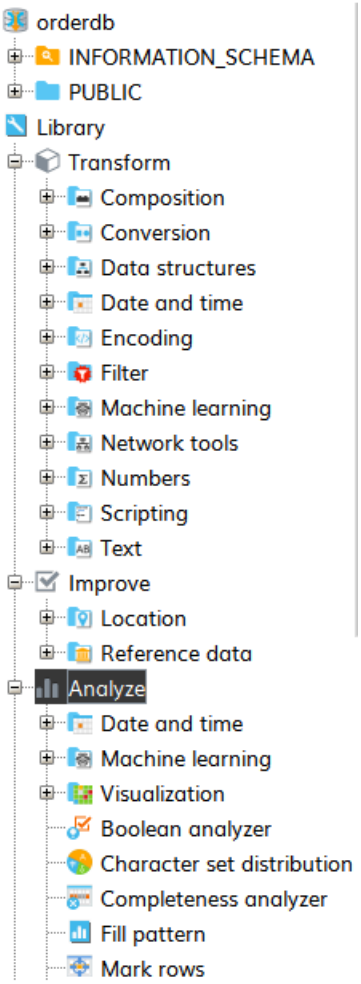
Analizadores: componentes más importantes, para que un job pueda ejecutarse, es necesario tener al menos un analizador, si no, DataCleaner aconsejará añadir uno básico que guardará el output en un fichero. Un analizador es componente que inspecciona los datos que recibe y genera un resultado o un informe.

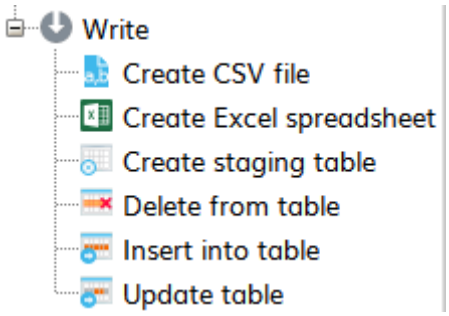
Transformadores: son componentes que modifican los datos antes de analizarlos, se utilizan para extraer valores, combinarlos, hacer lookups para luego añadir el resultado al flujo de datos del job. El resultado de un transformador es un set de columnas.

Filtros: componentes que dividen el flujo de procesamiento en un job. Un filtro tendrá una serie de resultados posibles y, dependiendo del resultado de un filtro, una fila en particular puede ser procesada por diferentes subflujos. Los filtros a menudo se usan simplemente para ignorar ciertas filas del análisis, por ejemplo. Valores nulos o valores fuera del rango de interés.

Cada uno de estos componentes se visualizará como un nodo en el grafo del job.

Árbol de componentes:

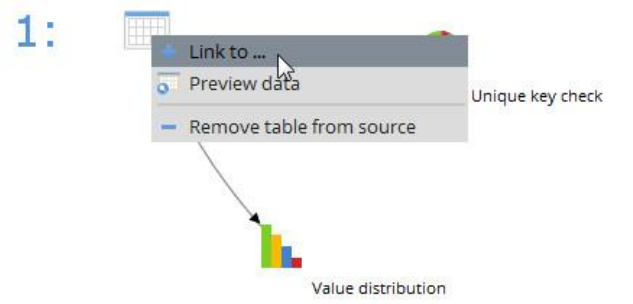




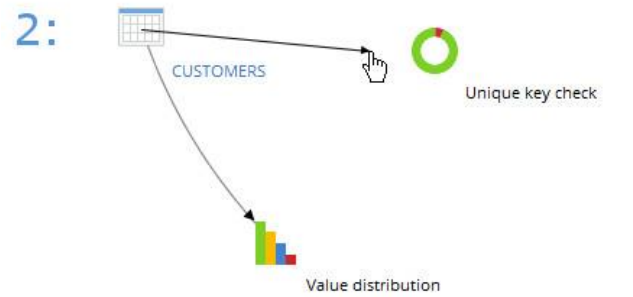
Podemos ver como hay 4 pestañas principales: Transform, Improve, Analyze y Write.

Los 3 tipos de componentes antes mencionados se reparten de esta forma: Transformadores y filtros están comprendidos en Transform e Improve, los analizadores están en la pestaña Analyze pero también están en la pestaña Write que nos permite formar outputs

Unión de nodos:

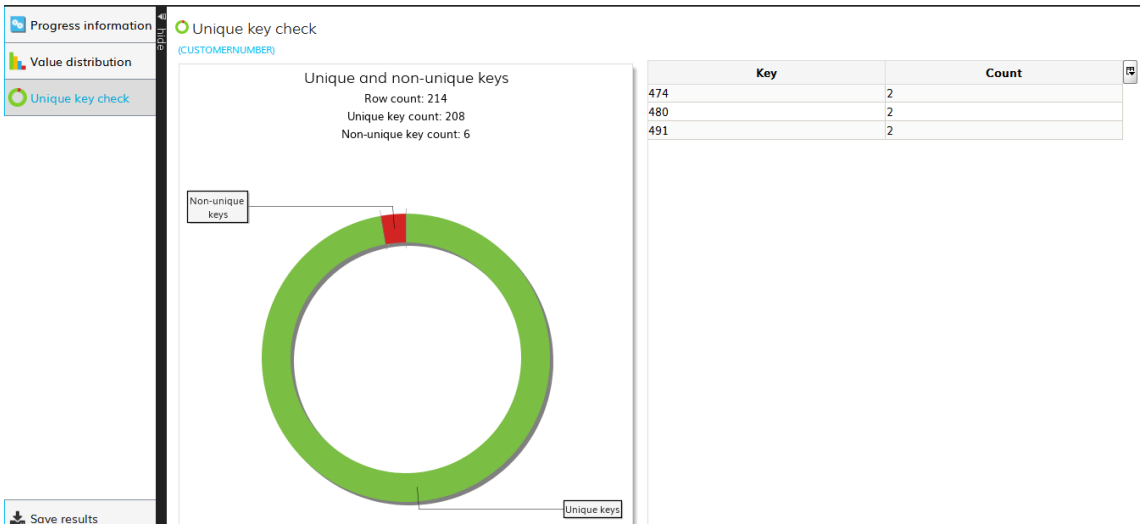


Vemos como en el paso 2 hay una fuente de datos, que es una tabla de una BD, la cual está conectada a un analizador que puede calcular la distribución de los valores de la tabla y visualizar gráficos.

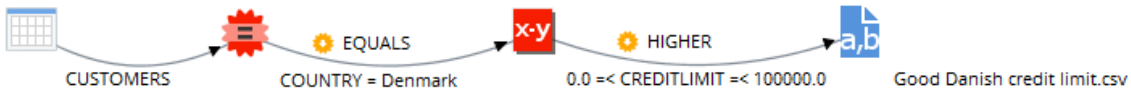


Una vez tengamos nuestro job montado, le daremos al botón Execute que está arriba a la derecha. Al

ejecutar el job, si tenemos analizadores, se nos abrirá una ventana que muestra los distintos analizadores empleados en el job y sus resultados, por ejemplo el resultado del job anterior sería:

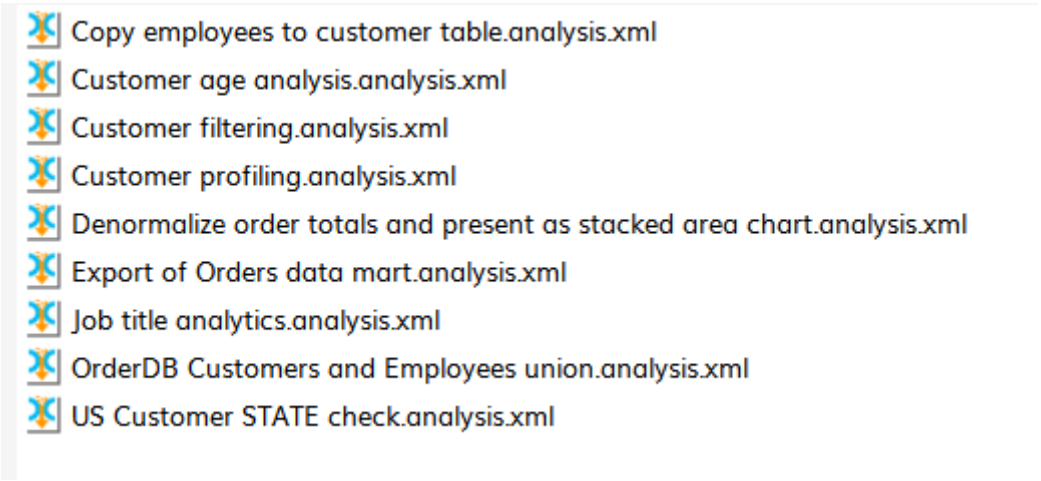


Aquí un ejemplo del uso de los filtros:

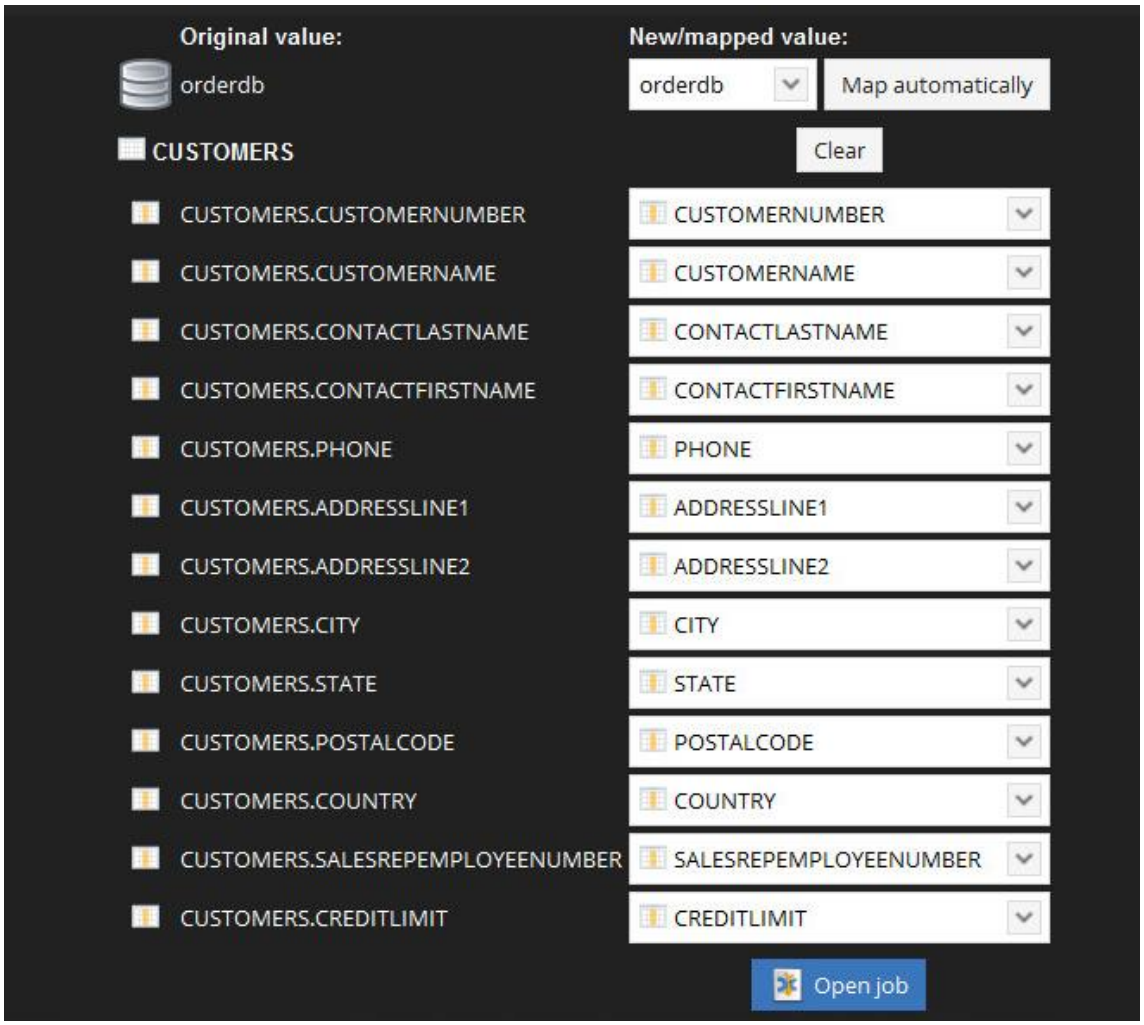


Los Jobs se guardan con formato xml de tal forma que podemos editarlos de forma manual, además DataCleaner porta plantillas (**template jobs**) que hacen referencia a operaciones comunes las cuales se pueden utilizar.

Con solo darle a open te muestra por defecto las plantillas predefinidas.



Además, es posible abrirla con el botón open as template que nos permitirá mapear los valores:

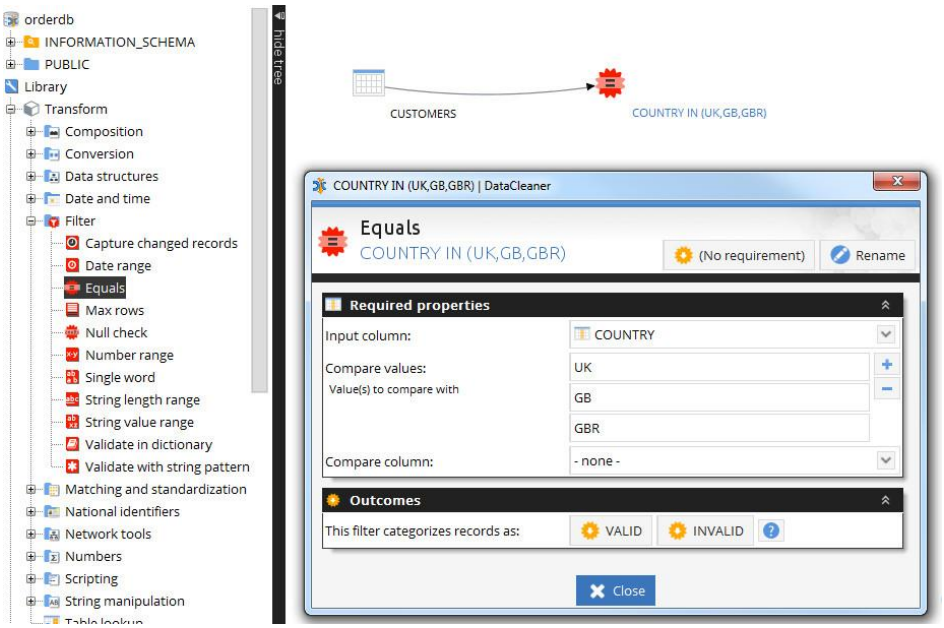


Algunos otros transformadores importantes son:

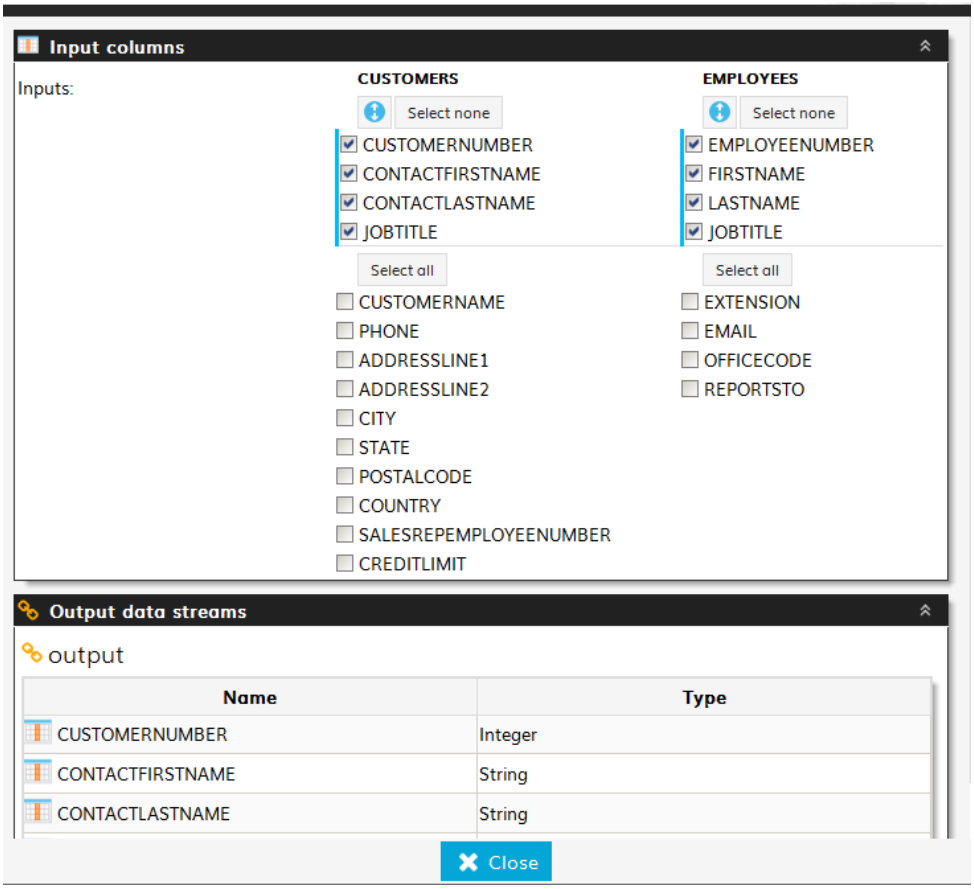
JavaScript: nos permite programar scripts en java, además contiene funciones predefinidas por DataCleaner.

Apply classifier y **Apply regression**: permiten aplicar un modelo de aprendizaje automático (machine learning) entrenado en los registros entrantes.

Equals: transformador de filtros que puedes hacer selección múltiple de los campos a comparar y además, puedes categorizar los registros como inválidos o válidos.



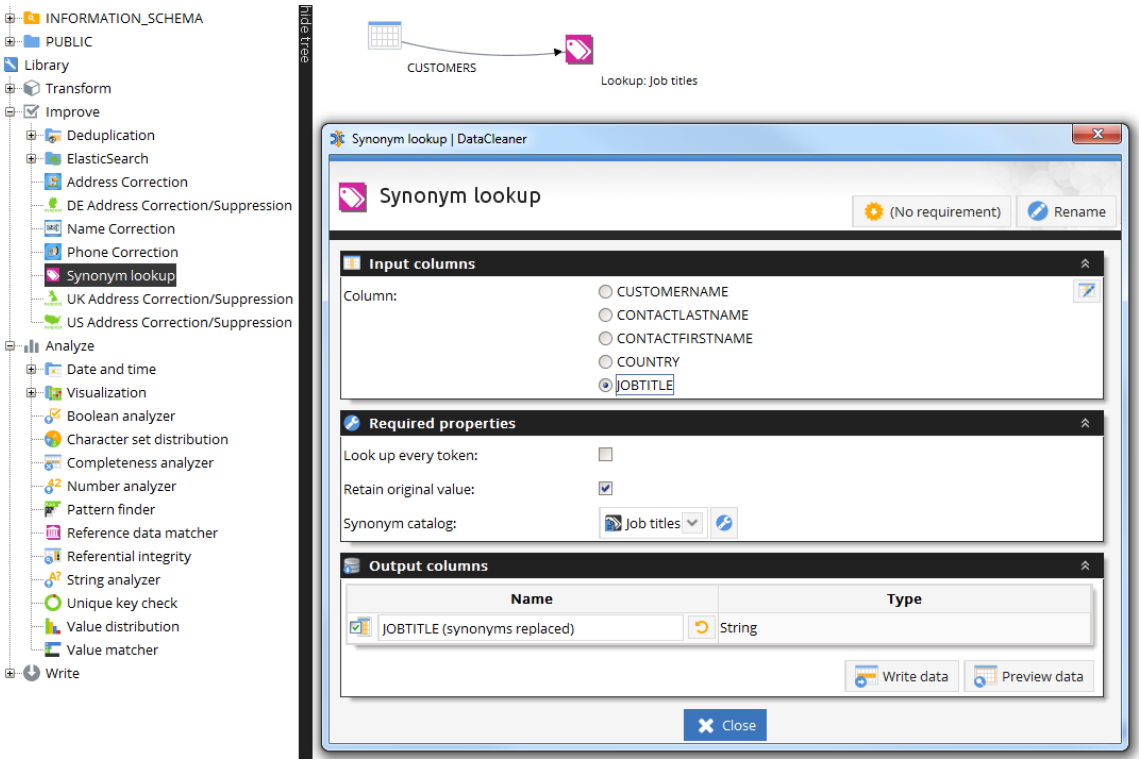
Union: permite combinar varios flujos de datos en uno.



3.5 Funciones de la pestaña Improve **y ejemplos de Data quality**

Diseñadas para asegurar la **calidad de los datos**, estas funciones no solo tienen la labor de analizar, sino que también, presentan una solución al posible problema. Un ejemplo de algunas de estas funciones sería: **Synonym lookup** o **Table lookup.**

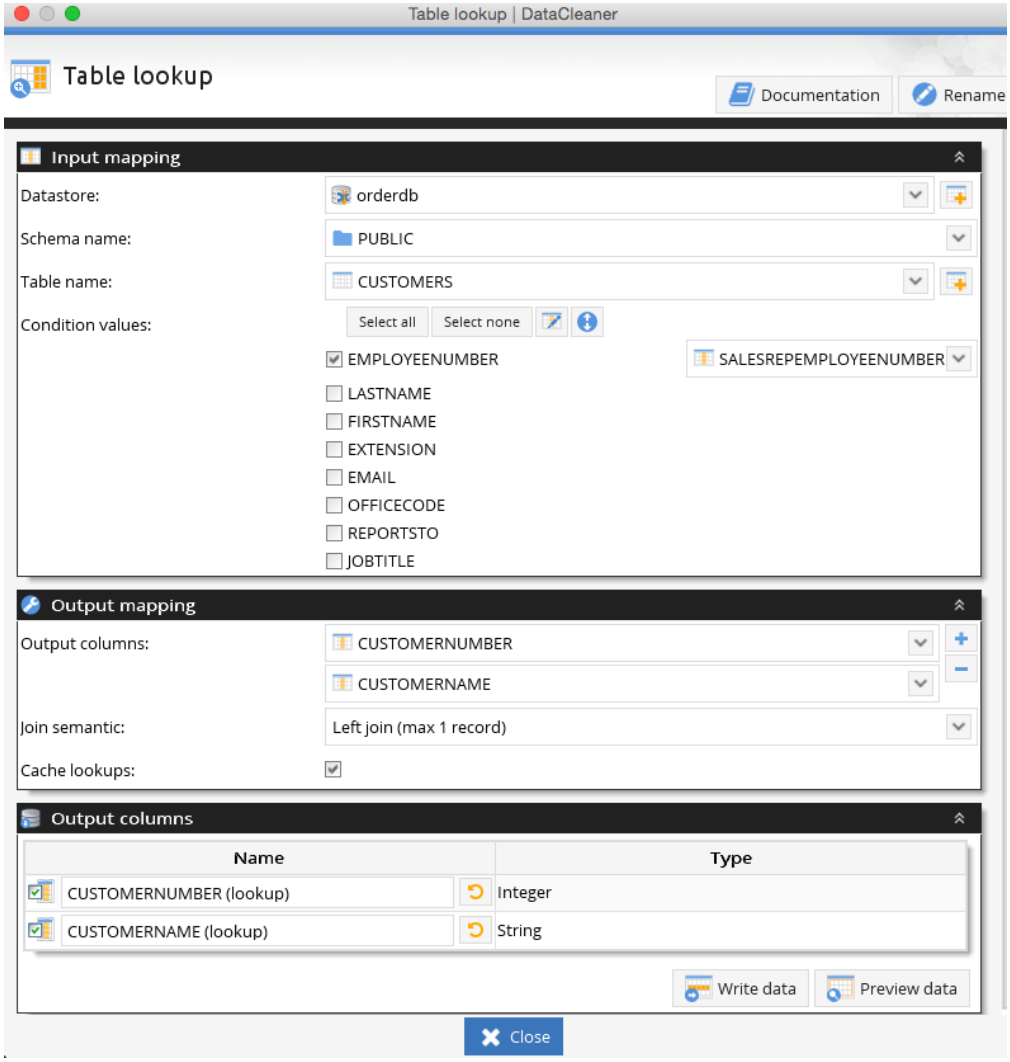
Synonym lookup: La transformación de búsqueda de sinónimos es una parte crítica de la capacidad de del usuario de DataCleaner para estandarizar y limpiar datos. Usando este componente puede buscar valores en un catálogo de sinónimos y reemplazarlo con su término maestro, si se encuentra que es un sinónimo. Un ejemplo:



En este ejemplo se ha seleccionado la columna a partir de la cual queremos realizar la búsqueda de sinónimos, la casilla **retain orignal value** hará que se devuelva el valor original en vez de un valor nulo si no se ha encontrado ningún sinónimo.

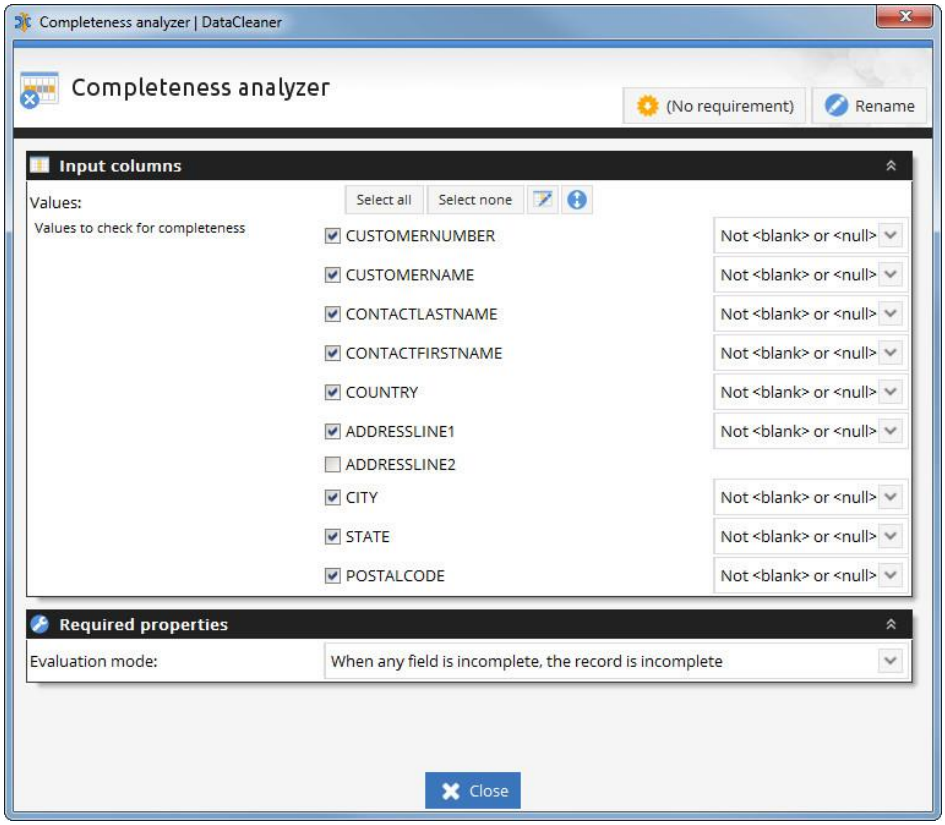
A continuación, veremos un ejmplo de Table lookup. Muy similar al que integra PDI.

Un ejemplo de **Table lookup**:



En la siguiente página se mostrará un analizador (de la pestaña analizadores) que asegura la **completitud de los datos**.

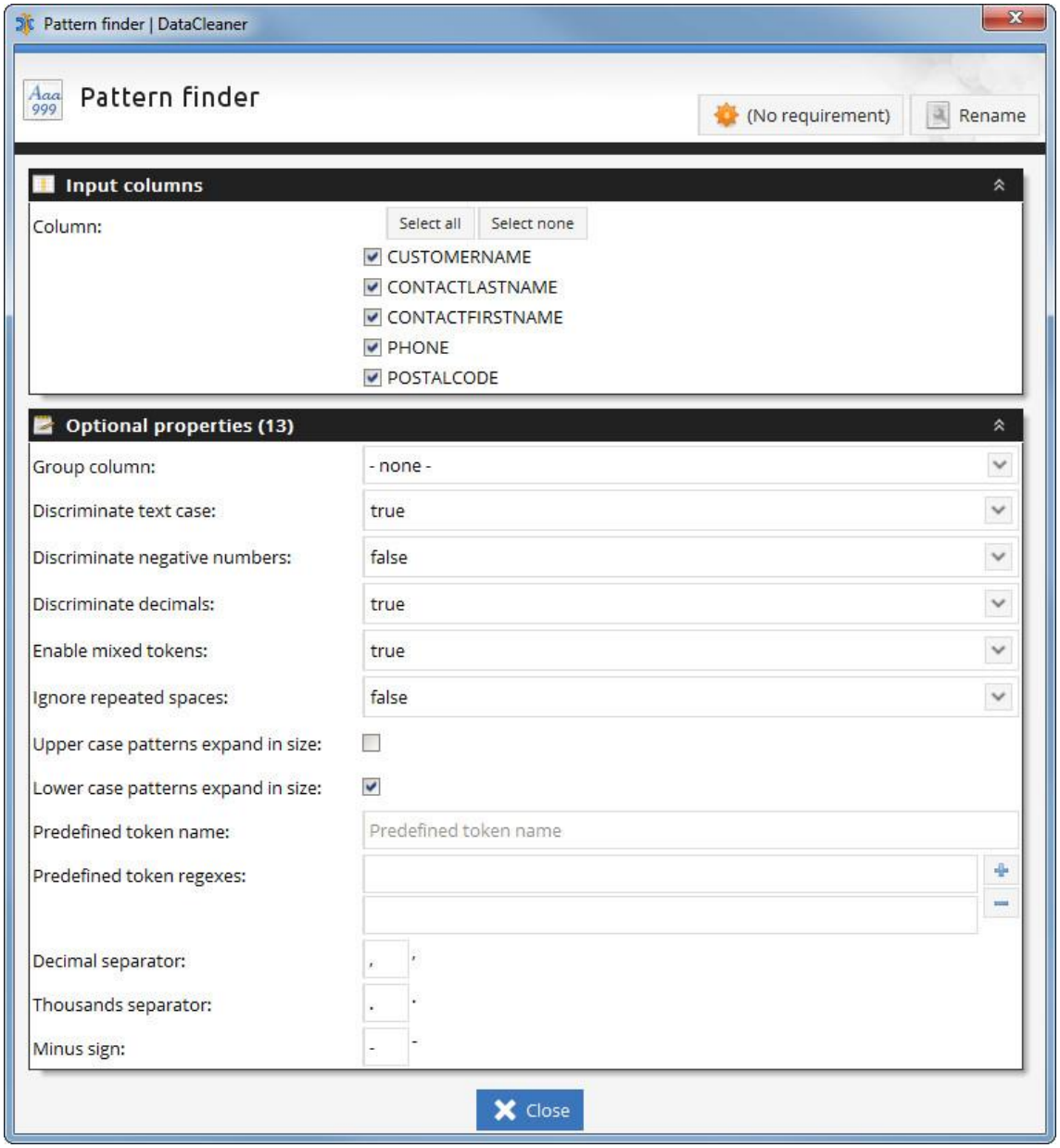
Para analizar la calidad de los datos tenemos varias funciones, una de ellas Completeness analyzer nos asegura la completitud de los datos, la cual es otra característica de **Data quality**.



Es posible especificar los campos de la tabla que no queremos que sean nulos y que en caso de que sí lo sean nos considere los campos según el Evaluation mode.

Otra función de análisis de calidad de datos y que, además es muy popular dentro de DataCleaner es **Pattern Finder**.

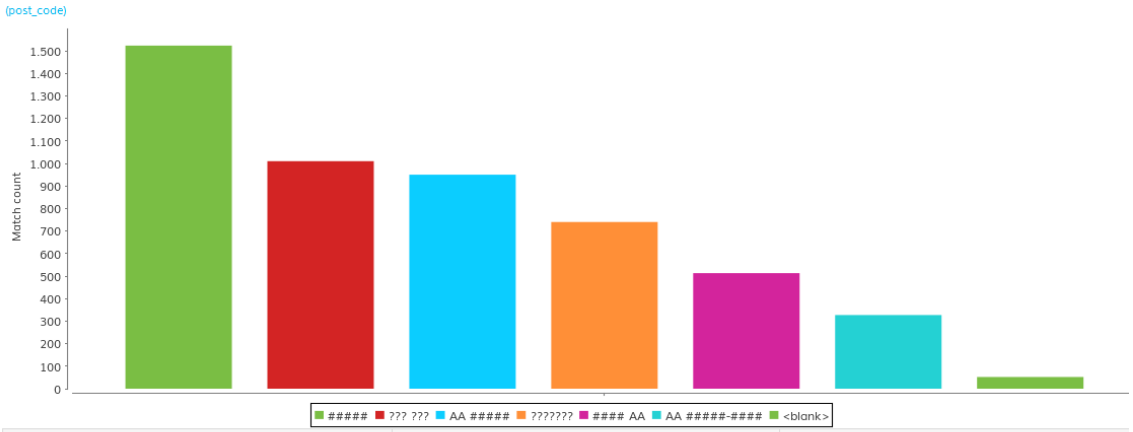
El buscador de patrones (Pattern finder) es uno de los analizadores más avanzados. A continuación, un ejemplo:



Por ejemplo, la propiedad pattern group permite definir una columna de grupo de patrones. Con una columna de grupo de patrones se puede separar los patrones identificados en grupos / grupos separados. Imaginemos, por ejemplo, que deseamos verificar si los números de teléfono de nuestros

clientes son consistentes. Si tenemos un cliente internacional, debe agruparse por una columna de país para asegurarse de que los patrones de teléfono identificados no coincidan con los patrones de teléfono de diferentes países.

Aquí un ejemplo del reporte resultante de la función Pattern Finder en este caso con código postal



	Match count	Sample
#####	1523	1530
??? ???	1010	CF5 5AE
AA #####	950	IA 52722
???????	740	SE207BW
#### AA	513	9645 GA
AA #####-####	327	CO 80022-3612
<blank>	52	<blank>

Como vemos en la tabla de abajo nos muestra los formatos diferentes en los cuales se presenta la columna código postal. Si clicamos en la flecha verde de las muestras nos hará **una select de todos los registros con ese formato encontrado**.

3.6 Conclusión sobre la herramienta

Aunque el principal enfoque de DataCleaner es el análisis, en muchos casos, te encontrarás con la necesidad de mejorar los datos (improve) mediante la aplicación de transformadores y filtros, es probable que estos datos (“limpios y mejorados”) sean útiles para operaciones futuras más allá del análisis. Habiendo investigando la interfaz, y realizado varios Jobs de ejemplo, podríamos ver esta herramienta como una mezcla entre PDI y PowerBI.