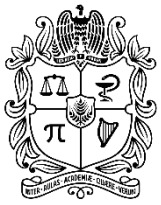


UNIVERSIDAD
NACIONAL
DE COLOMBIA

Comparación de herramientas ETL de código abierto

Jhoan Esteban Ruiz Borja

Universidad Nacional de Colombia
Facultad de Minas, Departamento de Ciencias de la Computación y de la Decisión
Medellín, Colombia
2018

Comparación de herramientas ETL de código abierto

Jhoan Esteban Ruiz Borja

Trabajo final presentado como requisito parcial para optar al título de:

Magister en Ingeniería de Sistemas

Directora:
Ph.D Claudia Jiménez Ramírez
Codirector:
Ph.D Juan David Velázquez Henao

Línea de Investigación:
Analítica

Universidad Nacional de Colombia
Facultad de Minas, Departamento de Ciencias de la Computación y de la Decisión
Medellín, Colombia
2018

Agradecimientos

Por el esfuerzo, apoyo y dedicación, presento mis más sinceros agradecimientos a la Profesora Asociada, Claudia Jiménez Ramírez y al Profesor Asociado, Juan David Velásquez Henao, de la Facultad de Minas de la Universidad Nacional de Colombia, Sede Medellín, quienes, como directora y codirector de la Tesis de Maestría, merecen mi respeto y admiración por su apoyo en la lectura cuidadosa, sugerencias y aporte presentados.

Adicionalmente, quiero ofrecer mis agradecimientos a mi familia por su incansable apoyo tanto económico, como aliciente y motivación para dar continuidad a mis estudios de posgrado en la Universidad Nacional de Colombia.

Por último, se agradece a todas las personas que de una u otra forma han participado o colaborado con su conocimiento en el desarrollo del trabajo investigativo. Reitero, a mi familia, por su apoyo y comprensión.

Resumen

El objetivo principal del presente trabajo final es la comparación de Pentaho Data Integration, Talend Data Integration y OpenRefine, tres herramientas de ETL (Extraction, Transformation and Load) de código abierto, con el propósito de promover la importancia que tiene en la actualidad el proceso ETL, y de facilitar no solo a los usuarios, sino también a empresas, que deseen aplicar algún proceso ya sea de extracción, transformación o carga de datos, mejorando el enfoque de inteligencia del negocio con estas herramientas adecuadas para el tratamiento de datos.

Se propuso describir varias herramientas de la actualidad, donde luego se describen los motivos de selección de las tres herramientas, prosiguiendo a la descripción detallada de las elegidas, para saber que capacidades poseen a la hora de realizar el proceso ETL, adicionalmente se dan los criterios de comparación, donde luego se presenta un ejemplo práctico, que permite comparar, para luego sugerir en qué casos puede ser más útil una herramienta con respecto a otra según sus características.

Por último, se presentan cuadros comparativos, donde se podrá resaltar las ventajas y desventajas de cada herramienta, junto con unas sugerencias que plantea qué herramienta utilizar para un caso dado, según la necesidad del usuario o empresa.

Abstract

The main objective of the present final work is the comparison of Pentaho Data Integration, Talend Data Integration and OpenRefine, three open source tools of ETL (Extraction, Transformation and Load), with the purpose of promoting the importance that the ETL process currently has, and to facilitate not only to the users, but also the companies, who wish to apply some process, involving Extraction, Transformation or Loading of data, to improve the Business Intelligence approach with these appropriate tools for data processing.

It was proposed to describe several current tools, then we describe the reasons for selecting the tools to be compared, continuing with the detailed description of the three tools chosen, to explore what capabilities they possess when carrying out the ETL process, additionally the comparison criteria are given, and then a practical example is presented, which allows comparing, and then suggesting in which cases a tool can be more useful with respect to another according to its characteristics.

Finally, comparative tables are presented, where the advantages and disadvantages of each tool can be highlighted, along with a brief guide that suggests which tool should be used for a specific case, according to the user's or company's need.

Contenido

Lista de figuras

Lista de tablas

1 INTRODUCCION

Hoy en día la cantidad enorme de datos generados de diferentes tipos de estructura o formatos crecen, convirtiéndose en un obstáculo para que las empresas puedan usar el enfoque de Inteligencia de Negocios en sus organizaciones. Por ese motivo, en la actualidad cada vez la competencia es más fuerte, por esto, la toma de decisiones, debe ser más acertada y oportuna para tener una mayor capacidad de lograr un objetivo determinado, o ser capaz de adaptarse a nuevas circunstancias. Todo esto implica, la necesidad de tener datos pre-procesados, de buena calidad y almacenados en un repositorio homogéneo, que sean útiles para realizar un análisis de datos, que apoye una mejor toma de decisiones. Para lograr ese objetivo, se debe tener un conocimiento base de herramientas ETL (Extraction, Transformation and Load), que ayuden a extraer, transformar y cargar datos, que luego se utilizan para realizar análisis, que sirve como apoyo a una mejor toma de decisiones. Por eso, hoy día la Extracción, Transformación y Carga de datos son fundamentales para cualquier empresa o institución, y es la etapa que mayor tiempo consume.

Existen varias herramientas en la actualidad, para realizar el proceso ETL, unas son comerciales (pagas) y otras son de código abierto (open source) que pueden ser utilizadas sin costo alguno. Sin embargo, los costos de algunas, ha llevado a que compañías que tienen bajos recursos o están iniciando como *Startup*s, busquen alternativas como las de código abierto que les permita realizar el proceso ETL, sin necesidad de realizar alta inversión. Luego, las compañías se preguntarán ¿Cuál es la herramienta más adecuada para realizar el proceso ETL? Para dar solución a esta pregunta, el presente trabajo tiene como propósito mostrar una comparación de tres herramientas ETL, con un caso práctico, implementado a partir de un conjunto de datos, donde se podrán evidenciar ventajas y desventajas de una herramienta con respecto a otra, que luego se mostrarán en un cuadro comparativo; finalizando con unas recomendaciones de los casos donde puede tener mejor rendimiento una herramienta con respecto a la otra.

1.1 Antecedentes

Existen en el mercado diversas herramientas destinadas a satisfacer el proceso de extracción (E), transformación (T) y carga (L) de datos desde múltiples fuentes, para reformatearlos, limpiarlos, y cargarlos en un repositorio para analizar y apoyar un proceso de toma de decisiones [1].

Se encontró información durante la investigación de varios artículos. El primero de ellos habla de la comparación de **“Pentaho Data Intregation”** y **“Talend Data Integration”** en su componente de código abierto, donde se comparó el proceso de instalación, la interfaz de usuario, manipulación de errores, ejecución de trabajo, entre otros, concluyendo que **“Pentaho”** obtuvo mejor rendimiento [2]. En el segundo artículo comparan 5 herramientas de ETL, entre las cuales están las dos mencionadas anteriormente; para la comparación se realizaron 22 pruebas y a cada prueba se le daba una calificación de 1 a 5, siendo 5 la mejor y 1 la no mejor, dando como resultado que **“Informatica”** obtuvo el mejor rendimiento [3]. En el tercer artículo comparan de nuevo a **“Pentaho Data Intregation”** y **“Talend Data Integration”**, procesando archivos de texto con 10.000, 100.000 y 5.000.000 de registros, de

los cuales se ejecutaron 4 pruebas para los archivos de 10.000 y 100.000 y se ejecutaron 2 pruebas para el archivo de 5 millones de registros, obteniendo como resultado que **"Talend"** obtuvo mejor rendimiento [4]. En la tesis comparan 7 herramientas de ETL del mercado, donde se compara la conectividad, las transformaciones complejas, velocidad de carga de datos, soporte técnico, interfaz de usuario y monitorización; terminando con el ejemplo práctico en **"Pentaho Data Integration"** [5]. En algunas páginas [6][7][8][9] comparan herramientas ETL comerciales y no comerciales, que exponen las características de cada herramienta, realizan un comparativo con un conjunto de datos, que luego resalta sus ventajas y desventajas, en común se encuentra que la mayoría se enfocan en la comparación de herramientas de ETL comerciales.

1.2 Planteamiento del problema

En la actualidad la herramienta de ETL se ha convertido en un punto clave en el proceso de desarrollo continuo dentro de las organizaciones. Es un componente indispensable que ayuda a integrar los datos provenientes de múltiples fuentes en un repositorio homogéneo, siendo una de las principales dificultades que se presenta en la fase de pre-procesamiento. En este sentido, la empresa posee en la actualidad una serie de procesos operativos que generan datos día a día, donde el problema surge en la correcta toma decisiones por parte de los entes directivos de la organización para cumplir con todas las metas propuestas. Es por eso que hay diseñado procesos de negocios que permita al ente llegar a tomar el camino acertado. Una de ellas es centralizar la información importante en un repositorio para analizar y emitir decisiones. Sin embargo, la falta de conocimiento de una herramienta de ETL, que le brinde una solución efectiva que minimice el tiempo y riesgo en cuanto al manejo de datos del proceso ETL, ya que el análisis de los datos generados por la operación es complejo. Este proceso cuando es realizado de forma manual y no automatizada genera en algunas ocasiones el surgimiento de riesgos que entorpecen el análisis, además, en la actualidad es un proceso demasiado lento y tedioso, especialmente cuando se tiene grandes cantidades de datos en distintas fuentes.
Es por todo esto, se hace necesario comparar tres herramientas ETL, para que de este modo se cree una propuesta que permita recomendar, en qué casos utilizar una herramienta con respecto a otra, obteniendo los mejores beneficios según su necesidad.

1.3 Justificación

Desde la experiencia en el trabajo actual, se enfrenta casos en los cuales las empresas, no poseen la información integrada en un solo repositorio, haciendo más complejo recopilar los datos, por toda la información que poseen en distintos formatos como archivos planos, archivos de Excel, entre otros. En algunos artículos [10][11][12][13], se entiende que el mundo empresarial se está volviendo cada vez más centrado en el consumidor, por estos motivos se ve una necesidad latente en las empresas de poder extraer, transformar y cargar datos de las diferentes fuentes heterogéneas, en un solo repositorio, que luego puedan utilizar para procesos analíticos, permitiendo obtener modelos y nuevo conocimiento que apoye a una mejor toma de decisiones. Motivo por el cual se realiza este trabajo, pretendiendo facilitar a las personas o compañías, por medio de unas recomendaciones, la elección de la herramienta de ETL más adecuada según su necesidad, y que sirva como herramienta óptima para hacer una tarea determinada.

1.4 Objetivos

1.4.1 Objetivo general

Describir y comparar Pentaho Data Integration, Talend Data Integration, OpenRefine, herramientas de proceso ETL de código abierto; para fortalecer el criterio de selección de la herramienta más adecuada según sus necesidades y propósito.

1.4.2 Objetivos específicos

1. Describir las características de Pentaho Data Integration, Talend Data Integration y OpenRefine.
2. Comparar Pentaho Data Integration, Talend Data Integration y OpenRefine en el proceso ETL, señalando ventajas y desventajas de cada una.
3. Presentar unas recomendaciones para seleccionar la herramienta más adecuada, según la necesidad y que permita más beneficios a la hora de utilizarla según sea el caso Extraer, Transformar o Cargar datos.

1.5 Alcance

Se pretende realizar una descripción y comparación de las herramientas, Pentaho Data Integration, Talend Data Integration y OpenRefine. También se realizará un ejemplo práctico con un conjunto de datos, para los procesos de extracción, transformación y carga de datos, donde se resaltarán ventajas y desventajas de las capacidades de cada herramienta, que luego se mostrarán en un cuadro comparativo, y por último se presenta una seria de recomendaciones de elección de herramienta según sea el caso de Extraer, Transformar o Cargar datos.

Se busca que sea un trabajo útil y de lectura rápida con buen entendimiento para las personas o empresas que buscan mejorar el criterio de selección de una de las herramientas mencionadas.

2 Fundamentos teóricos

En Inteligencia de Negocio (BI) se utiliza una gran cantidad de datos e información para el análisis, de modo que se espera obtener información importante. Este tipo de información se puede usar para apoyar el proceso de toma de decisiones. En la práctica, se necesita un proceso que recopile datos e información existentes en el almacén de datos. Este proceso de integración de datos se conoce como Extracción, Transformación y Carga (ETL). Hoy día, se han desarrollado muchas aplicaciones para llevar a cabo el proceso ETL, pero la selección de qué aplicaciones son más eficientes en tiempo, costo y rendimiento puede convertirse en un desafío. Por eso hoy día el proceso de extracción, transformación y carga – ETL (Extraction, Transformation and Load) es una de las actividades técnicas más críticas en el desarrollo de soluciones de Inteligencia de Negocios (BI). Hace parte del componente de integración y, de su implementación adecuada depende la integridad, uniformidad, consistencia y disponibilidad de los datos utilizados en el análisis. Su función es extraer, limpiar, transformar, resumir y formatear los datos que se almacenarán en una Bodega de Datos de la solución inteligencia de negocios.[14][15][16]

El proceso ETL puede dividirse en tres subprocesos o componentes: componente de extracción, componente de transformación y componente de carga. A continuación, se definen los conceptos que han sido considerados necesarios establecer previamente por su relación con el tema, que más adelante se complementan de una forma más detallada:

- **Proceso de Extracción, Transformación y carga (ETL):**
  - **Extracción:** Es el proceso de extraer datos con diferentes tipos de estructura, sistemas transaccionales, hojas de cálculo, archivos de texto, Web, XML, JSON, entre otros.
  - **Transformación:** Es el proceso de transformar, limpiar, filtrar y personalizar datos crudos.
  - **Carga:** Es el proceso de cargar datos formateados, estructurados a una Bodega de Datos de acuerdo a las necesidades.
- **Bodega de Datos (DW):** Según Inmon [17] una Bodega de Datos es una colección de datos orientados a temas, integrados, no-volátiles y variantes en el tiempo, organizados para soportar necesidades empresariales como el anális.
- **Inteligencia de Negocios (BI):** Es el conjunto de estrategias y herramientas utilizadas para la gestión y creación de conocimiento a partir del análisis de las Bodegas de Datos existentes dentro de la organización. Mediante la inteligencia se logra consolidar y analizar la información integrada con el proceso ETL y almacenada en una Bodega de Datos con razonable velocidad, detalle y precisión para ayudar a tomar mejores decisiones de negocio.[14][18]

2.1 Descubrimiento de conocimiento en bases de datos (KDD)

El proceso de extraer conocimiento a partir de grandes volúmenes de datos ha sido reconocido por muchos investigadores como un tópico de investigación clave en los sistemas de bases de datos, y por muchas compañías industriales como una importante área y una oportunidad para obtener mayores ganancias. Autores como Fayyad, Piatetsjy-Shapitri y Smith[19] lo definen como “El proceso no trivial de identificación de patrones válidos, novedosos, potencialmente útiles y fundamentalmente entendibles al usuario a partir de los datos”.

El Descubrimiento de Conocimiento en Base de datos (KDD, del inglés *Knowlegde Discovery in Database*) es básicamente un proceso automático en el que se combinan descubrimiento y análisis. El proceso consiste en extraer patrones en forma de reglas o funciones, a partir de los datos, para que el usuario los analice. Esta tarea implica generalmente pre-procesar los datos, hacer minería de datos (*Data Mining*) y presentar resultados. KDD se puede aplicar en diferentes dominios, por ejemplo, para determinar perfiles de clientes fraudulentos, para descubrir relaciones implícitas existentes entre síntomas y enfermedades, entre características técnicas y diagnósticos del estado de equipos y máquinas, para determinar perfiles de estudiantes “académicamente exitosos” en términos de sus características socioeconómicas y para determinar patrones de compra de los clientes en sus canastas de mercado, entre otros. [19][20][21]

2.1.1 Etapas del proceso KDD

El proceso KDD que se muestra en la figura 1 es interactivo e iterativo, involucra numerosos pasos con la intervención del usuario en la toma de muchas decisiones.

Se resume en las siguientes etapas:

- Selección.
- Pre-procesamiento/Limpieza.
- Transformación/Reducción.
- Minería de datos (Data Mining).
- Interpretación/Evaluación.

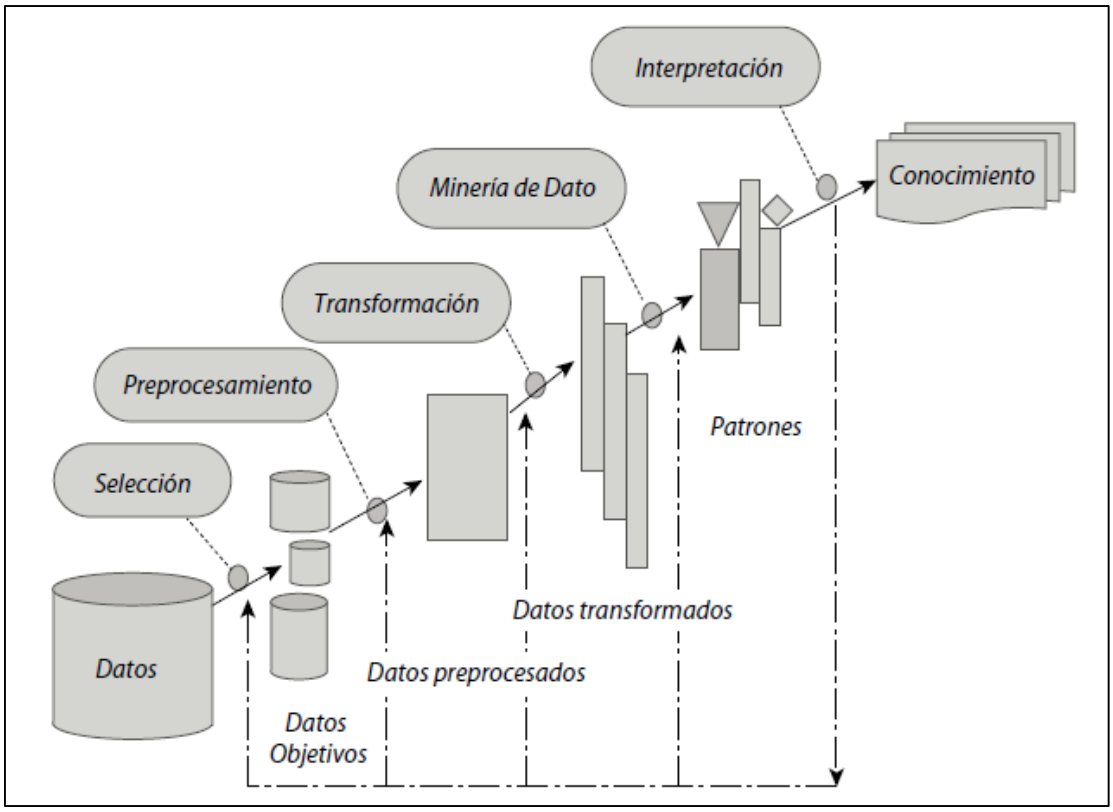


Figura 1: Etapas del proceso KDD. [19]

2.1.1.1 Etapa de selección

En la etapa de selección, una vez identificado el conocimiento relevante, prioritario y definidas las metas del proceso KDD, desde el punto de vista del usuario final, se crea un conjunto de datos objetivo, seleccionando todo el conjunto de datos o una muestra representativa de este, sobre el cual se realiza el proceso de descubrimiento. La selección de los datos varía de acuerdo con los objetivos del negocio. En esta etapa se ha de aplicar una componente del proceso ETL, como es la Extracción de datos de múltiples fuentes como datos estructurados, semi-estructurados y no estructurados. [19][20]

2.1.1.2 Etapa de pre-procesamiento/limpieza

En la etapa de pre-procesamiento/Limpieza (*Data Cleaning*) se analiza la calidad de datos, se aplican operaciones básicas como la remoción de datos ruidosos, se seleccionan estrategias para el manejo de datos desconocidos (*missing* y *empty*), datos nulos, datos duplicados y técnicas estadísticas para su reemplazo. En esta etapa, es de suma importancia la interacción con el usuario o analista.

Los datos ruidosos (*noisy data*) son valores que están significativamente fuera del rango de valores esperados; se deben principalmente a errores humanos, a cambios en el sistema, a información no disponible a tiempo y a fuentes heterogéneas de datos. Los datos desconocidos; *empty*, son aquellos a los cuales no les corresponde un valor en el mundo real y los *missing* son aquellos que tienen un valor que no fue capturado. Los datos nulos son datos desconocidos que son permitidos por los sistemas gestores de base de datos relacionales (SGBDR). En el proceso de limpieza todos estos valores se ignoran, se reemplazan por un valor por omisión, o por el valor más cercano, es decir, se usan métricas de tipo estadístico como media, moda, mínimo y máximo para reemplazarlos. En esta etapa se ha de aplicar una componente del proceso ETL, como es la Transformación de datos que provienen con múltiples estructuras, para unificarlos a un formato estándar que sea fácil de procesar a la hora de realizar análisis, se ha comprobado que en esta etapa se puede requerir aproximadamente un de 80% del proceso KDD. [19][20]

2.1.1.3 Etapa de transformación/reducción

En la etapa de transformación/reducción de datos, se buscan características útiles para representar los datos dependiendo de la meta del proceso. Se utilizan métodos de reducción de dimensiones o de transformación para disminuir el número efectivo de variables bajo consideración o para encontrar representaciones invariantes de los datos.
Los métodos de reducción de dimensiones pueden simplificar una tabla de una base de datos horizontal o verticalmente. La reducción horizontal implica la eliminación de tuplas idénticas como producto de la sustitución del valor de un atributo por otro de alto nivel, en una jerarquía definida de valores categóricos o por la discretización de valores continuos (por ejemplo, edad por un rango de edades). La reducción vertical implica la eliminación de atributos que son insignificantes o redundantes con respecto al problema, como la eliminación de llaves, la eliminación de columnas que dependen funcionalmente (por ejemplo, edad y fecha de nacimiento). Se utilizan técnicas de reducción como agregaciones, compresión de datos, histogramas, segmentación, discretización basada en entropía, muestreo, entre otras. En esta etapa se ha de aplicar dos componentes del proceso ETL, como es la Transformación de datos y la Carga de datos luego de haber sido formateados, se procede a cargar los datos en el repositorio que se desee. [19][20][21]

2.1.1.4 Etapa de minería de datos

El objetivo de la etapa de minería de datos es la búsqueda y descubrimiento de patrones insospechados y de interés, aplicando tareas de descubrimiento como clasificación, agrupación (clustering), patrones secuenciales y asociaciones, entre otras.

Las técnicas de minería de datos crean modelos que son predictivos o descriptivos. Los modelos predictivos pretenden estimar valores futuros o desconocidos de variables de interés, que se denominan variables objetivo, dependientes o clases, usando otras variables denominadas independientes o predictivas, como por ejemplo predecir para nuevos clientes si son buenos o malos basados en su estado civil, edad, género y profesión, o determinar para nuevos estudiantes si desertan o no en función de su zona de procedencia, facultad, estrato género, edad y promedio de notas. Entre las tareas predictivas están la clasificación y la regresión. Lo modelos descriptivos identifican patrones que explican o resumen los datos; sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos, como identificar grupos de personas con gustos similares o identificar patrones de compra de clientes en una determinada zona de la ciudad. Entre las tareas descriptivas se cuentan las reglas de asociación, los patrones secuenciales, los segmentos y las correlaciones.

Por lo tanto, la escogencia de un algoritmo de minería de datos incluye la selección de los métodos por aplicar en la búsqueda de patrones en los datos, así como la decisión sobre los modelos y los parámetros más apropiados, dependiendo del tipo de datos (categóricos, numéricos) por utilizar. [19][20][21]

2.1.1.5 Etapa de interpretación/evaluación de datos

En la etapa de interpretación/evaluación, se interpretan los patrones descubiertos y posiblemente se retorna a las anteriores etapas para posteriores iteraciones. Esta etapa puede incluir la visualización de los patrones extraídos, la remoción de los patrones redundantes o irrelevantes y la traducción de los patrones útiles en términos que sean entendibles para el usuario. Por otra parte, se consolida el conocimiento descubierto para incorporarlo en otro sistema para posteriores acciones o, simplemente, para documentarlo y reportarlo a las partes interesadas; también para verificar y resolver conflictos potenciales con el conocimiento previamente descubierto. [19]

2.2 Extracción, Transformación y Carga de datos

Las siglas ETL, proceden del inglés: Extraction, Transformation, and Load. Es decir, extracción, transformación y carga de los datos. Se trata, de uno de los procesos más importantes y decisivos para la correcta obtención de los resultados buscados, ya que, la calidad de los datos es fundamental.

Antes de explicar los procesos de extracción, transformación y carga de los datos, se debe entender, que dichos datos, proceden de las diferentes fuentes de las que se pueda obtener información relacionada con un negocio, estas fuentes son conocidas como **fuentes de información**.

Las fuentes de información de un negocio, son todas las aplicaciones, documentos, foros, libros, etc. Es decir, cualquier información susceptible de ser representada en los sistemas de información.

Algunas de las fuentes de información más comunes en las empresas son, las aplicaciones ERP[1], CRM,[2] las bases de datos, los documentos de texto o las hojas de cálculo. Además, se debe añadir cualquier otra información de la empresa, ya sea externa o interna.

Hay que destacar que el proceso de extracción, transformación y carga, consumen entre un 60% y 80% de un proyecto de Inteligencia de Negocio. Su principal objetivo, se encarga de una gran parte de la construcción de la Bodega de Datos del proyecto.[16][22]

2.2.1 Extracción de datos

Es el primer paso para la creación de una Bodega de Datos y es el proceso encargado de la recuperación física de los datos procedentes de las diversas fuentes de la información. Se dispone en este momento de los datos tal y como han sido recopilados, almacenados.
Los datos se pueden extraer de dos formas distintas, la primera, utilizando rutinas de programación que extraigan los datos de su origen o duplicando directamente los datos mediante los sistemas que proporcionan las propias bases de datos.

El segundo caso, consiste, en la utilización de herramientas especializadas para los procesos ETL. Así, se puede visualizar cada paso, siendo por tanto más fácil, la detección de errores.

El principal objetivo de la extracción es extraer solamente, los datos que serán útiles para responder preguntas que se plantean inicialmente en un proyecto, reduciendo así, los datos, y facilitando su entendimiento.

Se debe, por tanto, determinar qué datos, tienen una adecuada procedencia y calidad. Estos datos, se almacenan de forma temporal, ya que necesitan ser tratados antes de su uso. El gran problema con respecto al proceso de extracción de datos, se debe, a la gran diversidad de fuentes de la información que existen en la actualidad, puesto que, la gran mayoría de estas fuentes, tienen problemas de compatibilidad con las demás y presentan la información de formas muy distintas. [17][22][23]

2.2.2 Transformación de datos

Una vez realizada la extracción de los datos, el siguiente paso a realizar es la transformación y limpieza de estos, ya que los datos que se extraen no están depurados y pueden no ser válidos para la implementación de una solución de inteligencia de negocio adecuada. Por lo cual, el primer paso de este proceso consiste básicamente en la transformación de los datos extraídos para que sean útiles, se elimina la información duplicada, tratando la ausencia de valores, los valores contradictorios, estudiar los campos que se utilizan con más de un propósito.

[1] ERP: Enterprise Resource Planning
[2] CRM: Customer Relationship Management

Algunos procedimientos de limpieza de los datos son:

- División. Que consiste en separar los datos de un campo en campos más sencillos y unitarios, es decir, en la base de datos de donde estos se obtuvieron, puede darse el caso de que, en el campo de dirección de un cliente, se incluya el teléfono de este. Durante este proceso, se separará la dirección y el teléfono en dos campos independientes.
- Corrección de errores sintácticos y comprobación de la veracidad de los datos.
- Estandarización (Standardizing) o transformación de los datos estableciendo formatos predefinidos, que faciliten el entendimiento de los datos y su procesado.
- Buscar la relación entre los datos para simplificarlos. Se suelen crear tablas nuevas con dicha información facilitando su representación.

Cuando se han eliminado los posibles problemas en los datos, se procede con la transformación. Ésta se realiza como en la fase de extracción, siempre considerando el modelo de negocio elegido y el resultado al que se desea llegar. En este proceso se pueden incluir valores derivados y agregados para conseguir un mayor rendimiento, adaptando su estructura a la herramienta de destino.

El proceso de transformación se debe realizar cada vez que se realice una extracción de datos. Hay que destacar que la transformación de los datos permite establecer el nivel de granularidad[3] que se tendrá finalmente.[17][23]

2.2.3 Carga de datos

La carga de datos consiste en la incorporación de estos en la Bodega de Datos con el formato adecuado. Se debe comprobar si los datos subidos a la Bodega de Datos coinciden con los datos procedentes de la transformación realizada.

Se trata generalmente de un paso sencillo, pero es muy crítico, puesto que si la información cargada no es la deseada o se produce algún error durante el proceso de carga. Los datos, en este caso, pueden llevar a resultados equivocados y por ende a la toma de decisiones errónea.

En este paso se puede decidir la información que se desea cargar, por lo que se puede establecer, según las transformaciones realizadas, el nivel de granularidad o detalle de la información que tendremos en la Bodega de Datos.

Sin embargo, la situación más adecuada sería la elección del nivel de granularidad en la etapa de transformación de los datos y realizar la carga completa de estos, ya que así, se pueden comprobar de una forma más sencilla, si los datos cargados son o no los correctos y si se produjo un error.

[3] La granularidad representa el nivel de detalle al que se desea almacenar la información, por ejemplo (Año, Mes, Día, Horas, Minutos, Segundos; País, Departamento, Ciudad).

Al igual que el proceso de extracción y de transformación, este proceso hay que realizarlo cada vez que hay una actualización de los datos. [19][23] [24]

3 Selección de las herramientas de ETL a comparar

Los procesos ETL suelen ser muy complejos, sobre todo por la gran cantidad de información que existe en la actualidad, que proviene de diferentes fuentes de información con diferentes estructuras y que se intentan integrar en un entorno homogéneo.

En el mercado actual existen todo tipo de herramientas especializadas, que se diferencian según el formato en el que se encuentran los datos, el objetivo perseguido, y tecnología utilizada. Algunas de las herramientas que se pueden encontrar fácilmente en la actualidad se presentan en la siguiente tabla 3.1

Nombre	Descripción	URL
Informática PowerCenter	Es una plataforma de integración de datos empresariales que funciona como unidad para intercambio de datos, integración de datos en la nube, migración de datos, procesamiento de eventos complejos, enmascaramiento de datos, calidad de datos, replicación y sincronización de datos, virtualización de datos, gestión de datos maestros, y mensajería.[25]	**https://www.informatica.com/co/products/data-integration/powercenter.html** Con licencia
IBM Infosphere DataStage	Utiliza las características de un framework en paralelo de alto rendimiento y la notación gráfica para integrar datos en múltiples sistemas. Proporciona una potente plataforma escalable para la integración fácil y flexible de todo tipo de datos, incluidos big data en reposo (basado en Hadoop) o en movimiento (basado en secuencias), en plataformas distribuidas y *mainframe*. Gestiona la carga de trabajo y las reglas de negocio mediante la optimización del hardware. Está disponible en varias versiones, como Server Edition, Enterprise Edition y MVS Edition. Enterprise Edition presenta arquitectura de procesamiento paralela y trabajos paralelos. La edición de servidor representa principalmente los trabajos de servidor. La Edición MVS relacionada con trabajos de *mainframe*.[25]	**https://www.ibm.com/us-en/marketplace/datastage** Con licencia
Oracle Data Integrator (ODI)	Es una aplicación de software basada en ETL, que se utiliza para la transformación y fusión de datos o la integración de datos de alto volumen, alto rendimiento, hasta procesos basados en eventos y servicios de datos habilitados para SOA[4] mediante el agregado de paralelismo. El componente de arquitectura importante de ODI es el repositorio, que es la recopilación de todos los metadatos y se accede mediante el modo cliente-servidor o el modo de cliente ligero. Oracle Data Integrator también funciona en el área de preparación y transformación como soporte para otro software de Oracle.[24]	**http://www.oracle.com/technetwork/middleware/data-integrator/overview/index.html** Con licencia
Microsoft SQL Server Integration Services (SSIS)	Es un componente de la base de datos SQL Server que realiza la integración de datos en el entorno de Windows. La principal ventaja de SSIS es que no es costoso. Sin embargo, una desventaja significativa es que no funciona en un entorno que no sea Windows. SSIS se lanzó por primera vez con SQL Server 2005. SQL Server	**https://docs.microsoft.com/en-us/sql/integration-services/sql-server-integration-services?view=sql-server-2017**

[4] SOA: Service Oriented Architecture

	2008, 2012 también ha enriquecido el servicio de integración. En junio de 2016, se lanzó una nueva versión de SSIS. [26]	Con licencia
SAS ETL Studio	Ofrece una plataforma ETL integrada. SAS es uno de los líderes del mercado que combina aplicaciones de almacenamiento de datos e inteligencia para el proceso comercial tradicional. Proporciona la facilidad de extracción de datos multiproceso para acelerar la transferencia de datos y las operaciones relacionadas. SAS ayuda a reducir los datos duplicados o inexactos al proporcionar una interfaz de arrastrar y soltar, no necesaria de programación o SQL (lenguaje de consulta estructurado) para gestionar datos. SAS Data Integration Studio permite a los usuarios crear y editar rápidamente la integración de datos, capturar y gestionar automáticamente metadatos estandarizados desde cualquier fuente, visualizar y comprender fácilmente los metadatos empresariales. [25]	**https://www.sas.com/en_us/software/data-management.html** Con licencia
SAP Data Manager	SAP ha desarrollado un producto ETL con fuerte soporte para Hadoop[5], transmisión de datos y aprendizaje automático, que permite integrar grandes cantidades de información de forma sencilla.	**https://www.sap.com/latinamerica/products/data-services.html** Con licencia
Pentaho Data Integration	Integración de datos utilizando un enfoque basado en metadatos. Utiliza un entorno gráfico intuitivo. No hace falta escribir líneas de código para su utilización y dispone de plugins.	http://community.pentaho.com/projects/data-integration/ Con licencia y versión gratis
Talend Data Integration	Herramienta basada en Eclipse, para el proceso ETL que es uno de los procesos más importantes en la integración de datos.	https://es.talend.com/products/talend-open-studio/ Con licencia y versión gratis
OpenRefine	Es una poderosa herramienta para trabajar con datos desordenado, limpiándolos, y transformándolos a un formato deseado.	http://openrefine.org/ Versión gratis
Scriptella ETL Project	Herramienta de lanzamiento de script ETL. Utiliza sintaxis XML para sus scripts, los cuales pueden integrarse con scripts escritos en SQL, JavaScrot, JEXL, Velocity, etc. Algunas de las fuentes de entrada que acepta son LDAP, JDBC, XML, CSV, texto, entre otros.	http://scriptella.javaforge.com/ Versión gratis
Together	Se compone de varias herramientas separadas con funcionalidades ETL. Están desarrolladas en código Java y soporta la conexión con diferentes tipos de bases de datos (MSSQL, Oracle, DB2, QED, JDBC, MySQL,…) y acepta como entrada varios tipos de archivos (CSV, XML,…). Algunas herramientas son: TDC – Together Document Converter, TDT – Together Data Transformer, TXE – Together XML Extractor.	http://www.together.at/download Versión gratis
Xineo XIL	Define un lenguaje XML para transformar fuentes de datos basadas en registros en archivos XML. Soporta JDBC y estructuras de texto.	http://software.xineo.net/xil.jspx
CloverETL Community Edition	Es una herramienta muy gráfica que permite varios tipos de transformaciones, así como diversos tipos de entrada y salida de datos, como son los procedentes de las BBDD MySQL, PostgreSQL, SQLite, MSSQL, Oracle, Sysbase y Derby, archivos CSV, XML, etc. Cuenta con versiones de pago que permiten muchas más opciones (clasificación, clusters).	http://www.cloveretl.com/products/community-edition Versión gratis

[5] Hadoop es un framework que soporta aplicaciones distribuidas bajo licencia libre.

Apatar	Usa interfaz gráfica de trabajo mediante la cual se puede hacer el filtrado, la validación y la planificación de los datos. Los conectores incluyen MySQL, PostgreSQL, Oracle, MSSQL, Sybase, FTP, HTTP, SalesForce.com, SugarCRM, Compiere ERP, CRM Goldmine, XML, archivos planos, WebDAV, Buzzsaw, LDAP, Amazon y Flickr. No se requiere. Todos los metadatos se guardan en archivos XML.	http://www.apatar.com/ Versión gratis
Jaspersoft ETL	Herramienta basada en Eclipse, para el proceso ETL que es uno de los procesos más importantes en la integración de datos. Incluye flujos y procesa diferentes tipos de archivos. Fácil de desplegar.	http://community.jaspersoft.com/project/jaspersoft-etl Versión gratis
Data Pipeline	Transforma datos y los procesa. Puede leer y escribir archivos de tipo CSV, Excel, JDBC, JSON.	http://northconcepts.com/data-pipeline/ Versión gratis
KETL	Está basado en java. Incluye gestión de Jobs y alertas. Es capaz de gestionar varios hilos a la vez. Los Jobs están definidos en XML.	http://www.ketl.org/ Versión gratis

Tabla 3-1: Herramientas ETL empresariales y de código abierto [22]

Se puede observar que existen varias herramientas ETL, que pueden facilitar los procesos de extracción, transformación y carga de datos. Las más interesantes son las herramientas que utilizan interfaces gráficas (GUI), ya que son más fáciles e intuitivas de utilizar.

Las herramientas mencionadas, usan técnicas muy diferentes a la hora de manipular los datos, utilizan distintos tipos de elementos de entrada y salida, y al ser varias herramientas de código abierto, algunas fueron desarrolladas por necesidades propias del programador. Por todo ello hay que tener claro, los tipos de datos, de dónde serán extraídos y qué formato deben de tener estos para poder usarse posteriormente en el resto de componentes BI.

Figura 2: Cuadrante Mágico de Gartner de Herramientas de ETL [27]

Según el cuadrante mágico de Gartner del 2018 en la figura 2, aparecen ocho de las herramientas de ETL mencionadas en la tabla 3.1, teniendo a seis de ellas como líderes y a dos que buscan mejorar para llegar a ser líderes. Sin embargo, solo dos de ellas hasta el momento comparten con la comunidad, sus componentes de código abierto como los son “Talend” que se encuentra entre los líderes y “Pentaho” que no está entre los líderes.

Según varios blogs [28][29][30][31][32][8][6] y artículos [3][4][2], [25], [33], [34] en años anteriores las herramientas más mencionadas y comparadas de código abierto son “Pentaho Data Integration (PDI)” y “Talend Data Integration (TDI)”, recientemente hasta año 2018, no se ha encontrado comparaciones, y tampoco hay comparaciones entre herramientas con “OpenRefine (OR)” que tiene un esquema de trabajo muy diferente al de las mencionadas anteriormente. Aunque existen muchas más herramientas, cada una con sus capacidades que la diferencia de las demás. Se aclara que “Jaspersoft ETL” [35] es una herramienta muy similar a “Talend Data Integration” en su forma de trabajar manejando las misma interfaz de usuario basada en Eclipse, y muchos de sus componentes de trabajo funcionan igual, como se ilustra en la figura 3 y 4.

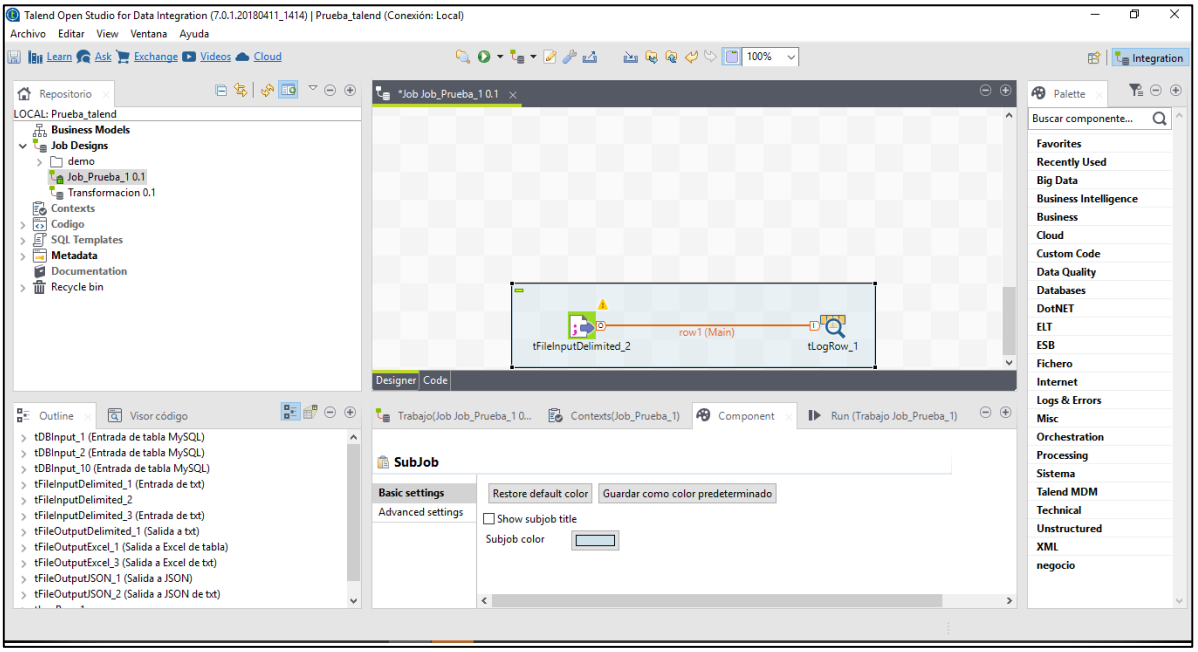


Figura 3: Interfaz gráfica de Talend Data Integration
Fuente: Elaboración Propia

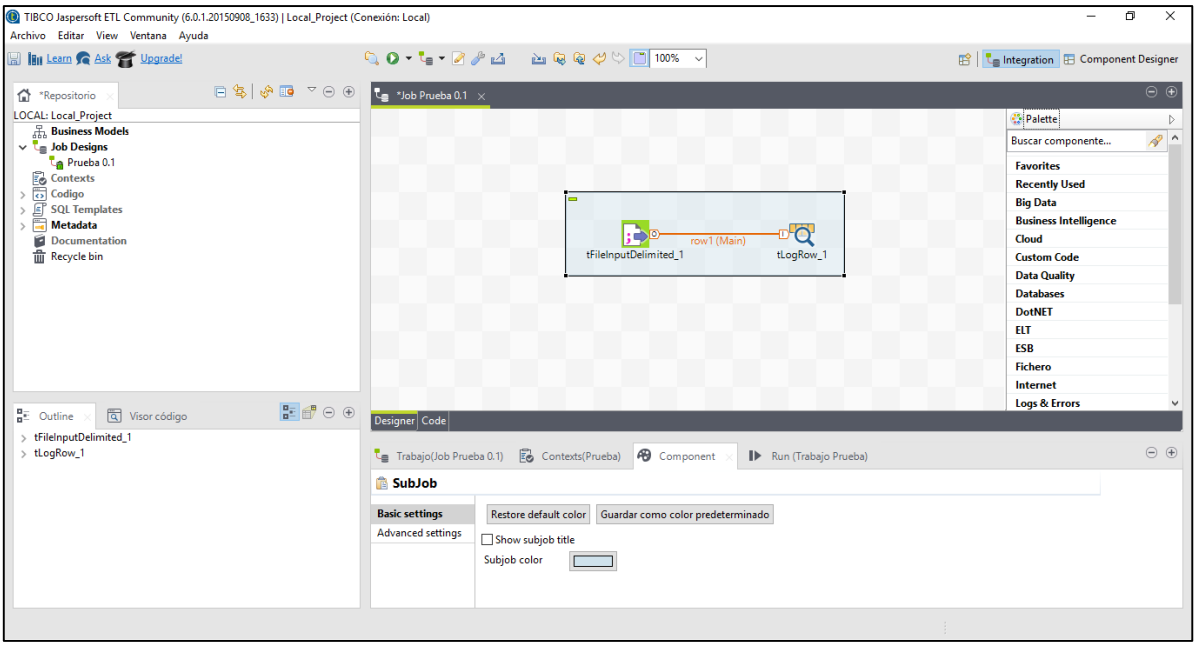


Figura 4: Interfaz gráfica de Jaspersoft ETL
Fuente: Elaboración Propia

Todas las herramientas de ETL citadas anteriormente son buenas opciones, pero dependiendo de las necesidades, el entorno de trabajo, las posibilidades o la estrategia de negocio, algunas pueden encajar mejor que otras en la empresa. Dependiendo de las motivaciones, se pueden establecer preferencias. Como es habitual, la selección de una herramienta siempre va a depender de las necesidades de la empresa, el perfil de los usuarios y entre otras consideraciones. Los criterios utilizados para elegir las herramientas “Pentaho Data Integration”, “Talend Data Integration” y “OpenRefine” fueron las capacidades que poseen en extraer, transformar y cargar datos, que hacen que sean de gran utilidad. Para ello se tuvo en cuenta, de qué fuentes las herramientas pueden extraer datos (bases de datos, la nube, archivos planos, archivos de Excel, XML, big data, entre otros) y la facilidad con que lo hacen, siendo este un punto en consideración ya que si una herramienta que permita poder extraer datos de estas distintas fuentes será de gran utilidad para el cumplimiento del objetivo buscado, otra consideración es la transformación de datos, es decir (filtrar, unir, limpiar, entre otros) y la facilidad con que lo hacen, permitiendo procesar toda la información correspondiente del proceso de extracción, y por último está el proceso de carga de datos, en el cual se tuvo en cuenta, las distintas fuentes en las que se puede cargar los datos es decir (bases de datos, la nube, archivos planos, archivos de Excel, XML, big data, entre otros) y la facilidad con que lo hacen.

A continuación, se describirán las herramientas Pentaho Data Integration, Talend Data Integration y OpenRefine que permiten realizar proceso ETL y que serán las que se comparan en este trabajo.

3.1 Pentaho Data Integration (PDI)

Pentaho Data Integration es una suite de Pentaho, que fue fundado en el 2001 por Richard Daley, bajo la organización “Hitachi Vantara”, que es la compañía creadora, PDI es hecha en Java con un conjunto de herramientas responsables de los procesos de Extracción,

Transformación y Carga. Es una herramienta de código abierto, multiplataforma (Window, Linux, Mac) desde la versión 2.2.0 que fue liberado al dominio público bajo la licencia (Licencia Pública General Reducida, o LGPL), utiliza la componente “Spoon” que es el diseñador gráfico de transformaciones y trabajos del sistema, también cuenta con una versión comercial. PDI no sólo sirve como una herramienta ETL, sino que también se utiliza para otros propósitos, como la migración de datos entre aplicaciones o bases de datos, la exportación de datos a bases de datos a archivos planos, entre otras. La versión de código abierto de PDI se puede descargar en el siguiente enlace:

[https://sourceforge.net/projects/pentaho/files/Data%20Integration/]

Entre las funciones de PDI encontramos:

- Conexión a múltiples bases de datos (Oracle, MySQL, PostgreSQL, entre otras).
- Extraer datos de múltiples fuentes (csv, txt, bases de datos, xls, xlsx, entre otros).
- Transformar datos.
- Cargar datos a múltiples formatos (txt, csv, xls, xlsx, entre otros).
- Consulta SQL.
- Integrar datos provenientes de múltiples fuentes en un archivo deseado.
- Manipulación de datos.
- Validación de datos y manipulación de errores.
- Guarda el flujo de trabajo.
- Visualización de datos.
- Explorador de base de datos.
- Flujos de procesos.
- Grandes volúmenes de datos (Big Data).
- Estadística.
- Bodega de Datos.
- Envío de mensajes.

Pentaho Data Integration cuenta en sus características principales la de manejar trabajos llamada **Job** y transformaciones llamada **Transformation**. Una transformación (**Transformation**) es una red de tareas lógicas llamadas pasos (**steps**), las transformaciones son esencialmente flujos de datos, cuyos nombres de archivos de transformación tiene una extensión (.ktr). Los trabajos (**Job**) son modelos similares al concepto de proceso, donde un proceso es un conjunto sencillo o complejo de tareas con el objetivo de realizar una acción determinada. En los trabajos se puede utilizar pasos específicos (que son diferentes a los disponibles en las transformaciones) como recibir un archivo, mandar un email, ejecutar un comando, entre otros. Además, se puede ejecutar una o varias transformaciones que se hayan diseñado. [36][37][38][39]

3.2 Talend Data Integration (TDI)

Talend Open Studio for Data Integration es una suite de Talend, que fue fundado en el 2005 por Mike Tuche, bajo la organización “Enterprise Software” que es la compañía creadora, TDI es una herramienta de código abierto hecha en Java, especializada en la integración de grandes volúmenes de datos (Big Data), que proporciona la extracción, transformación y carga de datos en la nube, gestión de datos maestros, calidad de datos, preparación de datos y la

integración de aplicaciones empresariales de software y servicios, que también cuenta con una versión comercial. La versión de código abierto deTDI se puede descargar en el siguiente enlace:

[https://es.talend.com/products/talend-open-studio/]

Entre las funciones de TDI encontramos:

- Importar metadatos.
- Integración de metadatos.
- Sincronizar metadatos.
- Genera script SQL.
- Limpiar datos.
- Conexión a múltiples bases de datos.
- Extraer datos de múltiples fuentes.
- Transformar datos.
- Cargar datos a múltiples formatos.
- Calidad de datos.
- Consulta SQL.
- Manipulación de datos.
- Validación de datos y manipulación de errores.
- Guarda el flujo de trabajo.
- Visualización de datos.
- Explorador de base de datos.
- Flujos de procesos.
- Manipulación de cadenas.

Talend Data Integration ofrece un diseñador visual (basado en Eclipse RCP) que permite definir todo el flujo de transformaciones en base de componentes predefinidos. Este diseñador proporciona una vista gráfica de los procesos, con componentes que se arrastran, que permiten transformaciones. TDI posee en su estructura dos grandes tipos de flujos que le permiten realizar su trabajo que son el **Main** y el **Iterate**, el Main es un tipo de flujo transaccional que marca un inicio y un final donde devuelve toda la información, mientras que el Iterate marca un inicio, manda un registro y marca un final, que permite ejecutar tantas veces como datos devuelva el componente anterior. [40][41][42][43]

3.3 OpenRefine (OR)

OpenRefine (anteriormente Google Refine) es una poderosa herramienta para trabajar con datos desordenados: limpiándolos, transformándolos de un formato a otro, extendiéndolos como servicios web y datos externos. OpenRefine se puede descargar del siguiente enlace:

[http://openrefine.org/]

Entre las funciones de OR encontramos:

- Importar datos en varios formatos como (TSV, CSV, SV, xls, xlsx, JSON, XML y RDF).
- Datos en documentos de Google.

- Explorar conjuntos de datos en cuestión de segundos.
- Aplicar transformaciones básicas y avanzadas a celdas.
- Tratar con celdas que contienen valores múltiples.
- Crear enlaces instantáneos entre conjuntos de datos.
- Filtra y dividir datos fácilmente con expresiones regulares.
- Realizar operaciones de datos avanzadas con el lenguaje general de expresión de precisión.

Google Refine encuentra su raíz en la solución Freebase Gridworks desarrollada por Metaweb Technologies, en mayo de 2010. Desde su primera versión, Freebase Gridworks era un proyecto de código abierto. Inicialmente fue una herramienta diseñada para la base de datos y la comunidad de Freebase para la limpieza, reconciliación y carga de datos. Este vínculo histórico con Freebase aún está presente en Google Refine, ya que la solución admite la comunicación con la base de datos de Freebase.

En julio de 2010, Google adquirió Metaweb, y por extensión, Freebase y Gridworks. Freebase Gridworks pasó a llamarse Google Refine y el código y la documentación se movieron a una instancia de code.google.com. El recientemente renombrado Google Refine continuó siendo un proyecto de código abierto para la limpieza de datos. Durante el período 2010-2012, con el apoyo de los ingenieros de Google y la comunidad, se realizaron tres actualizaciones de Google Refine (2.0, 2.1 y 2.5).

Actualmente hay interés de OpenRefine por parte de varias comunidades. Bibliotecarios, periodistas y analistas de datos han estado utilizando OpenRefine para limpiar y reconciliar sus datos por medio de la interfaz de fácil uso que ayudó a miles de usuarios no técnicos a tomar el control de sus datos, ayudando en el mundo de Big Data y OpenRefine permite reducir la barrera técnica para participar y capacitar a más personas para el análisis y procesamiento de datos. [44][45]

OpenRefine cuenta con una versión 3.0 Beta, y una versión 2.8 estable la cual se pueden instalar en diferentes sistemas operativos como Windows, Mac y Linux. También cuenta con algunas operaciones usando una de las bibliotecas existentes, esas bibliotecas están utilizando la API de OpenRefine, en lenguajes de programación como Python, R, Ruby, PHP y Javascript, para potencializar las capacidades de la herramienta. OpenRefine tiene como apoyo de aprendizaje un libro llamado “Using OpenRefine” de Ruben Verborgh y Max De Wilde lanzado en el 2013, que permite un acercamiento a la herramienta mostrando la forma de instalar junto con sus capacidades. Actualmente la versión 2.8 estable no se puede conectar a bases de datos, pero en la versión 3.0 Beta ya se le agregó una componente de bases de datos que le permite establecer una conexión con las bases de datos (MySQL, PostgreSQL y MariaDB), lo que permite potencializar el proceso de ETL en OpenRefine [46]

Nota: Freebase es una base de conocimiento colaborativo compuesta por metadatos creados principalmente por miembros de su comunidad. Y Metaweb es la compañía propietaria, que luego pasó a ser comprada por Google.

4 Criterios de comparación más comunes de herramientas de ETL

Según varios artículos, blogs y tesis [2][45][47][9][48][49][50], algunos de los criterios de comparación más comunes se mencionan a continuación.

- **Facilidad de uso:** En este criterio se compara cuál de las herramientas es más fácil y sencilla de aprender a utilizar sin gran esfuerzo.
- **Visualización:** En este criterio se compara cuál de las herramientas permite visualizar los datos de una forma fácil y sencilla.
- **Conectividad:** En este criterio se compara las conexiones a distintos tipos de estructura de datos que permite cada herramienta.
- **Velocidad de lectura y escritura de datos:** En este criterio se compara cuál de las herramientas es más rápida en lectura y escritura de datos tanto en la extracción, transformación y carga de datos.
- **Tiempo que tarda en ejecución el proceso:** En este criterio se compara el tiempo que tarda en ejecución las herramientas en el proceso de extracción, transformación y carga de datos.
- **Implementación**: Se compara los requisitos mínimos necesarios para utilizar las herramientas de ETL.

A continuación, se describe cada criterio de comparación de una forma más detallada.

Facilidad de uso

Cuando se habla de facilidad de uso, se refiere a la facilidad de poder instalar la herramienta sin necesidad de realizar configuraciones previas, se habla de la facilidad de poder realizar un proceso ETL, sin conocimientos previos que lleven mucho esfuerzo a la hora de utilizar la herramienta, se habla de la disponibilidad de una interfaz gráfica intuitiva, sencilla de entender y aprender.

Visualización

Cuando se habla de visualización, se refiere a la facilidad con la que se pueden ver los datos antes y después de un proceso ETL, sin tener que esperar a que el proceso culmine al cargar los datos para poder visualizar los resultados.

Conectividad

Habilidad para conectar con un amplio rango de tipos de estructura de datos, que incluyen bases de datos relacionales y no relacionales, varios formatos de archivos, XML, aplicaciones ERP, CRM, SCM, emails, sitios web, la nube, entre otros.

Velocidad de lectura y escritura de datos

Cuando se habla de velocidad de lectura y escritura de datos se habla de la habilidad para leer y escribir gran cantidad de datos en un tiempo determinado (usualmente Segundos) de cualquier tipo de estructura de datos como las mencionadas en conectividad.

Tiempo que tarda en ejecución del proceso

Se refiere al tiempo requerido para realizar algún proceso ETL, es decir si se desea extraer datos de cualquier tipo de estructura, qué tiempo tardaría en ejecutar este proceso, siendo el mismo caso para la transformación y para la carga de datos.

Implementación

Se refiere a los sistemas operativos en los que se puede implementar, y qué capacidad mínima se requiere de CPU, RAM y lenguaje de programación necesario para su implementación.

A continuación, se mostrará una tabla con la importancia que se le da a cada criterio de comparación, entre baja, media y alta. [9][48]

Criterio de comparación	Importancia
Facilidad de uso	Alta
Visualización	Media
Conectividad	Alta
Velocidad de lectura y escritura de datos	Media
Tiempo que tarda en ejecución del proceso	Media
Implementación	Baja

Tabla 4-1: Criterios de comparación con nivel de importancia

El proceso de comparación de las herramientas de ETL, se realizará en un portátil, en el cual se harán ejemplos de extracción, transformación y carga de datos. El portátil tiene las siguientes características:

- 16 GB de RAM
- Disco duro de una Tera (No es de estado sólido)
- Un procesador Core i7 de séptima generación
- Sistema operático Windows 10 Pro

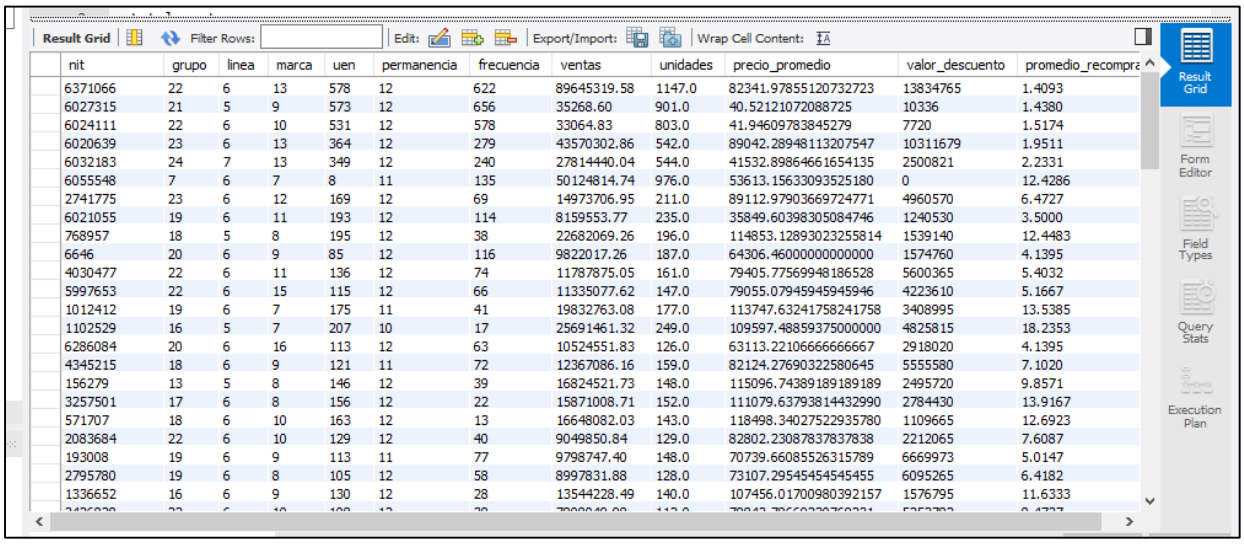
5 Comparación de herramientas en el proceso de extracción de datos (E)

La extracción de datos consiste en realizar una copia de los datos mediante una selección de lo requerido. Se pueden extraer tablas completas, algunos campos, archivos completos, algunos registros, entre otros, dependiendo de los requerimientos establecidos. Estas extracciones se hacen sobre las fuentes de información con las que se cuente y que hayan sido seleccionadas para alimentar la bodega de datos.

Como PDI, TDI y OR pueden establecer conexiones con múltiples fuentes de datos como (Bases de datos, archivos planos, archivos de Excel, JSON, XML, entre otros), para la extracción de datos, el ejemplo práctico se realizará extrayendo datos de una tabla de la base de datos MySQL, y de un archivo plano, por motivos de comparar con dos conjuntos de datos, en formatos diferentes, las herramientas, es decir comparar herramientas en extraer datos de una tabla de base de datos y luego comparar herramientas en extraer datos de un archivo plano, y también porque los datos no solo provienen de una fuente, donde se busca comparar la velocidad de lectura y escritura de datos, tiempo que tarda en ejecución el proceso y facilidad de extracción de datos.

5.1 Extracción de datos de una tabla de MySQL

Para comparación de PDI, TDI y OR, en cuando a la velocidad de lectura y escritura de datos, tiempo que tarda en ejecución del proceso y facilidad de extracción de datos, se usará una tabla de MySQL con 50.000 registros y 61 campos, qué son de retail (figura 5), cuyos datos son de fuente propia de una empresa. A continuación, se describe como extraer datos de una tabla de MySQL con PDI, TDI y OR.



nit	grupo	linea	marca	uen	permanencia	frecuencia	ventas	unidades	precio_promedio	valor_descuento	promedio_recompra
6371066	22	6	13	578	12	622	89645319.58	1147.0	82341.97855120732723	13834765	1.4093
6027315	21	5	9	573	12	656	35268.60	901.0	40.52121072088725	10336	1.4380
6024111	22	6	10	531	12	578	33064.83	803.0	41.94609783845279	7720	1.5174
6020639	23	6	13	364	12	279	43570302.86	542.0	89042.28948113207547	10311679	1.9511
6032183	24	7	13	349	12	240	27814440.04	544.0	41532.89864661654135	2500821	2.2331
6055548	7	6	7	8	11	135	50124814.74	976.0	53613.15633093525180	0	12.4286
2741775	23	6	12	169	12	69	14973706.95	211.0	89112.97903669724771	4960570	6.4727
6021055	19	6	11	193	12	114	8159553.77	235.0	35849.60398305084746	1240530	3.5000
768957	18	5	8	195	12	38	22682069.26	196.0	114853.12893023255814	1539140	12.4483
6646	20	6	9	85	12	116	9822017.26	187.0	64306.46000000000000	1574760	4.1395
4030477	22	6	11	136	12	74	11787875.05	161.0	79405.77569948186528	5600365	5.4032
5997653	22	6	15	115	12	66	11335077.62	147.0	79055.07945945945946	4223610	5.1667
1012412	19	6	7	175	11	41	19832763.08	177.0	113747.63241758241758	3408995	13.5385
1102529	16	5	7	207	10	17	25691461.32	249.0	109597.48859375000000	4825815	18.2353
6286084	20	6	16	113	12	63	10524551.83	126.0	63113.22106666666667	2918020	4.1395
4345215	18	6	9	121	11	72	12367086.16	159.0	82124.27690322580645	5555580	7.1020
156279	13	5	8	146	12	39	16824521.73	148.0	115096.74389189189189	2495720	9.8571
3257501	17	6	8	156	12	22	15871008.71	152.0	111079.63793814432990	2784430	13.9167
571707	18	6	10	163	12	13	16648082.03	143.0	118498.34027522935780	1109665	12.6923
2083684	22	6	10	129	12	40	9049850.84	129.0	82802.23087837837838	2212065	7.6087
193008	19	6	9	113	11	77	9798747.40	148.0	70739.66085526315789	6669973	5.0147
2795780	19	6	8	105	12	58	8997831.88	128.0	73107.29545454545455	6095265	6.4182
1336652	16	6	9	130	12	28	13544228.49	140.0	107456.01700980392157	1576795	11.6333

Figura 5: Datos de la tabla de MySQL
Fuente: Elaboración Propia

5.1.1 Extracción de datos de una tabla MySQL con PDI

Para la extracción de datos de una tabla de MySQL con PDI, primero se debe crear una transformación. Se debe ir a "**File**", luego a "**New**", y dar clic en la opción "**Transformation**", como se observa en la figura 6.

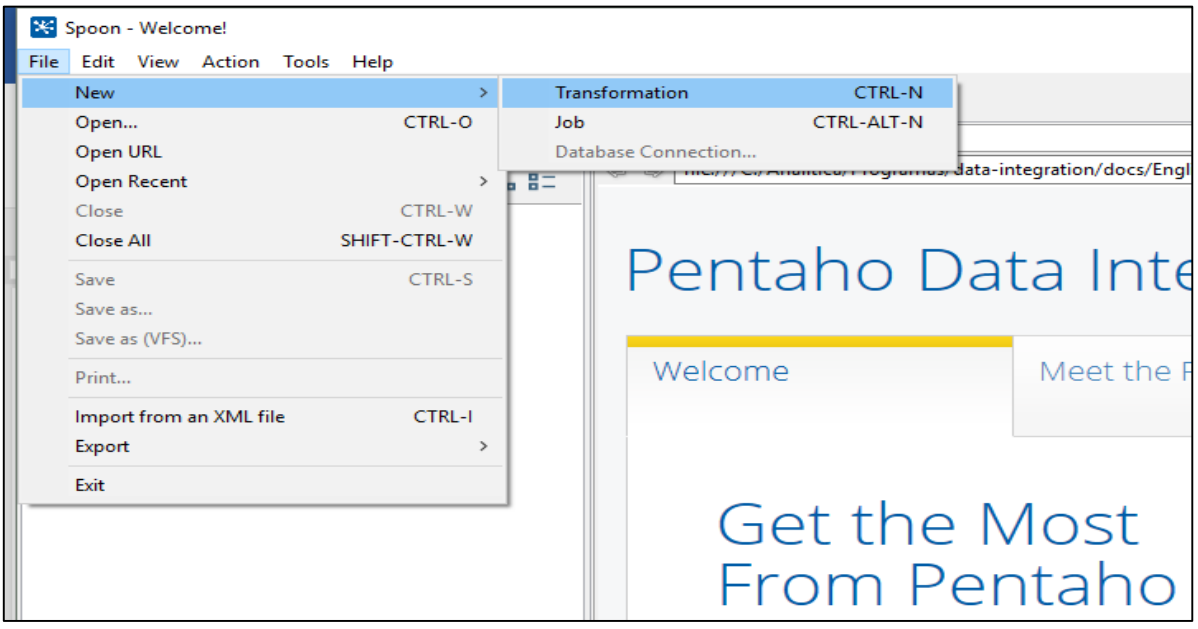


Figura 6: Creación de una transformación para la extracción de datos con PDI
Fuente: Elaboración Propia

Luego de haber creado la transformación se va al contenedor “**Input**”, que contiene los componentes de entrada de datos en PDI, y se arrastra el componente “**Table input**”, que sirve para crear una conexión con la base de datos MySQL, como se ilustra en la figura 7.

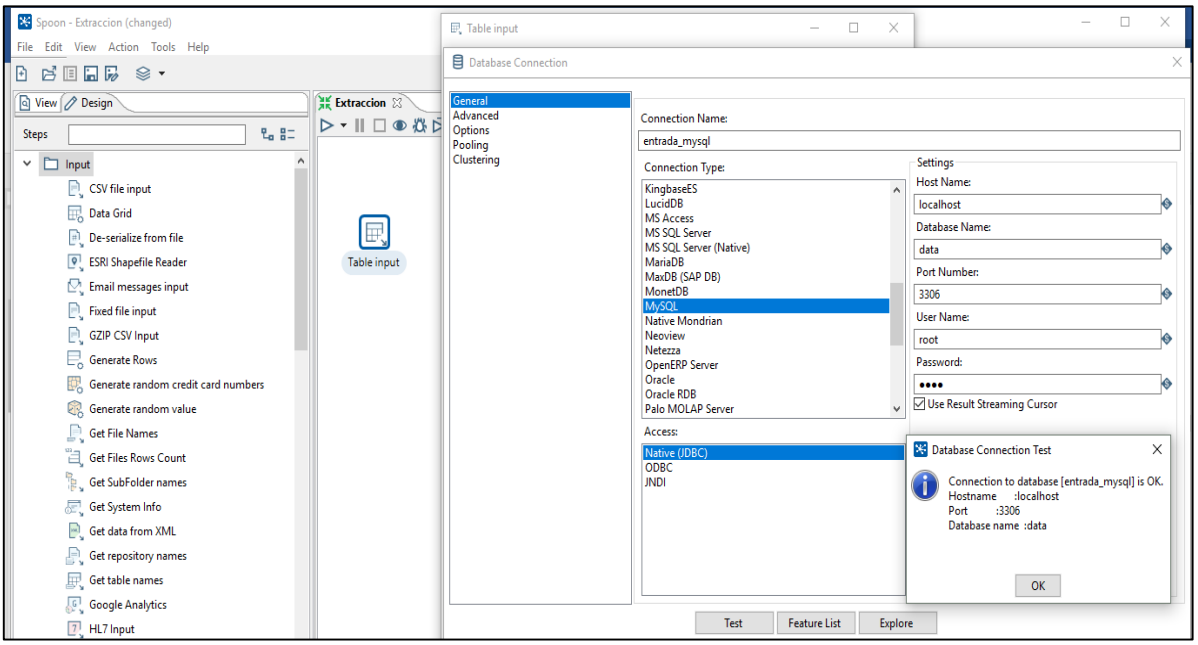


Figura 7: Conexión a la base de datos MySQL con PDI
Fuente: Elaboración Propia

Después se debe indicar a dónde se van a cargar los datos, en este caso se realiza la carga en un archivo plano, se va al contenedor “**Output**” que contiene los componentes de salida de datos de PDI, y se arrastra el componente “**Text file output**”, que sirve para cargar los datos en un archivo plano, como se puede ver en la figura 8.

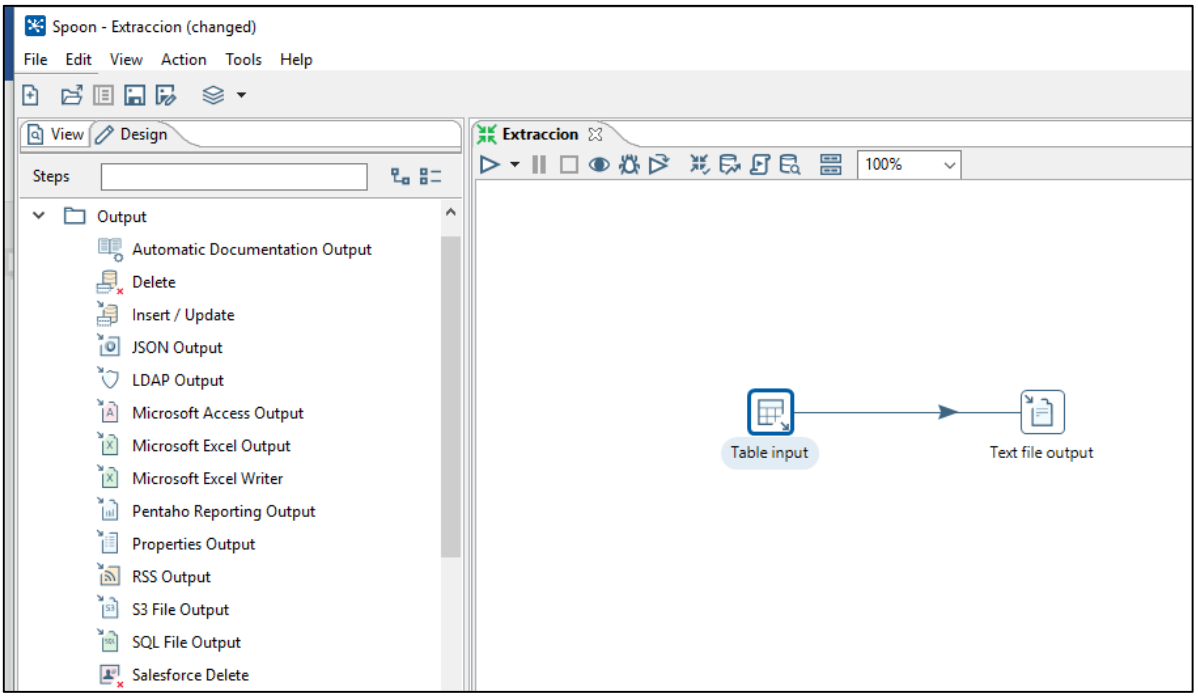


Figura 8: Componente de salida de datos con PDI
Fuente: Elaboración Propia

Posteriormente se renombran los componentes “**Table input**” como “**Entrada de MySQL**” y “**Text file output**” como “**Salida a txt**”, y se prosigue con la consulta SQL, que sirve para extraer los datos que se pueden visualizar con el botón “**Preview**”, como se ilustra en la figura 9.

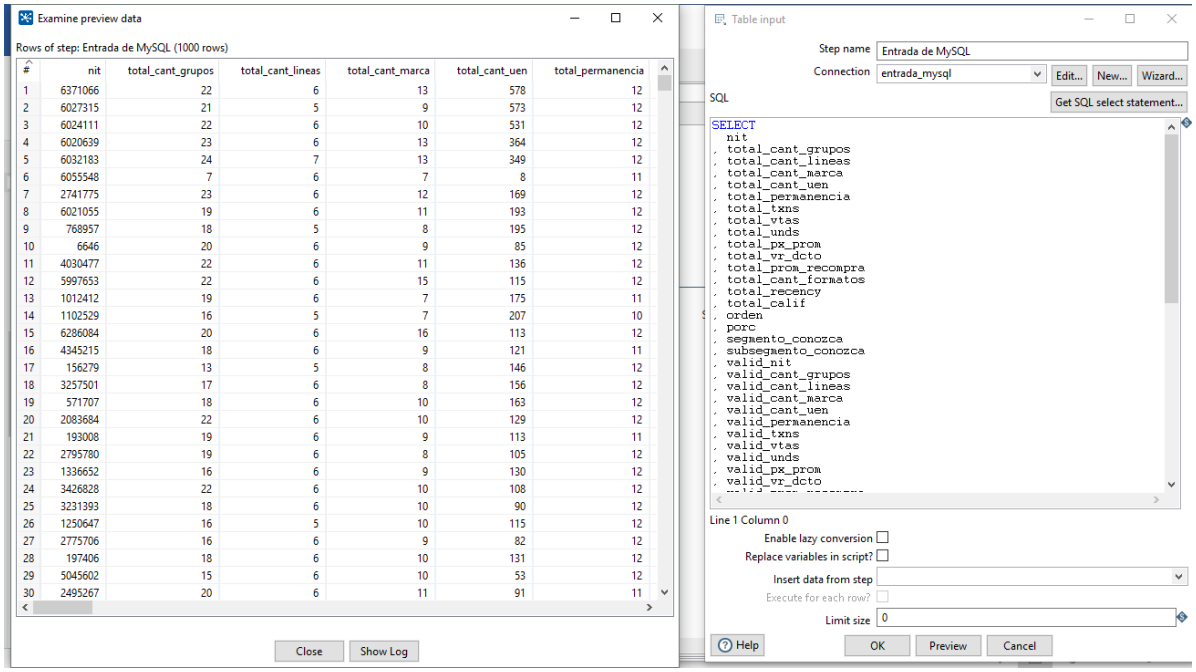


Figura 9: Consulta SQL para la extracción de datos de la tabla de MySQL con PDI
Fuente: Elaboración Propia

Continuando con el proceso, se debe configurar el componente de salida de datos “**Salida a txt**”, indicando la ruta donde se van a guardar los datos, como se puede observar en la figura 10.

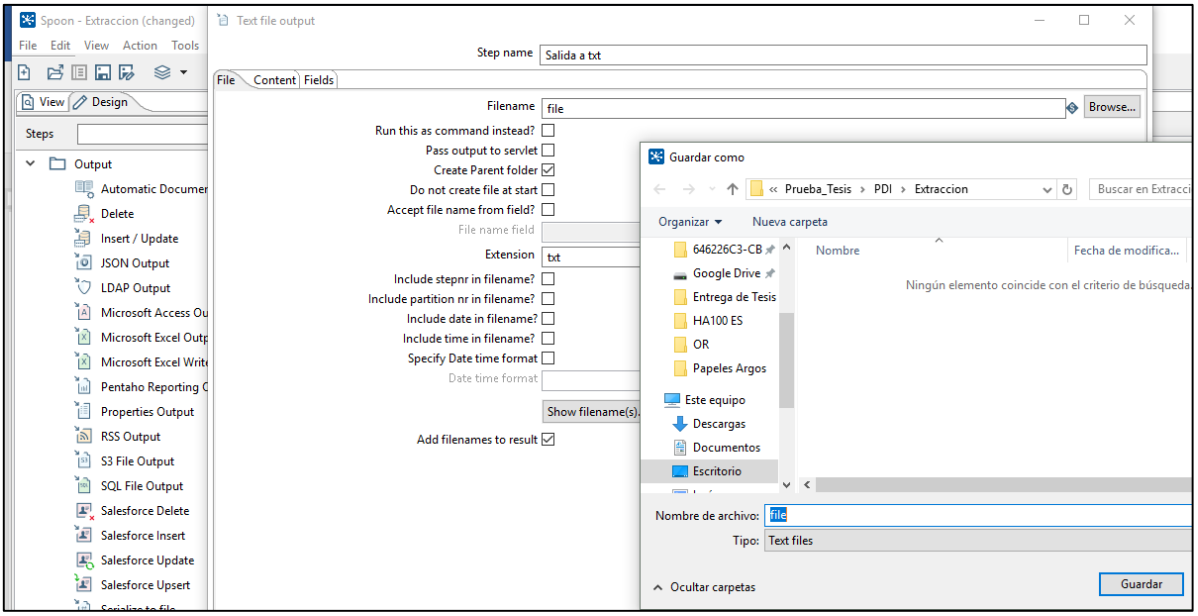


Figura 10: Configuración de salida de datos al archivo plano con PDI
Fuente: Elaboración Propia

Por último, después de tener el proceso de extracción de datos, en la figura 11 se observan los resultados luego de haber ejecutado PDI.

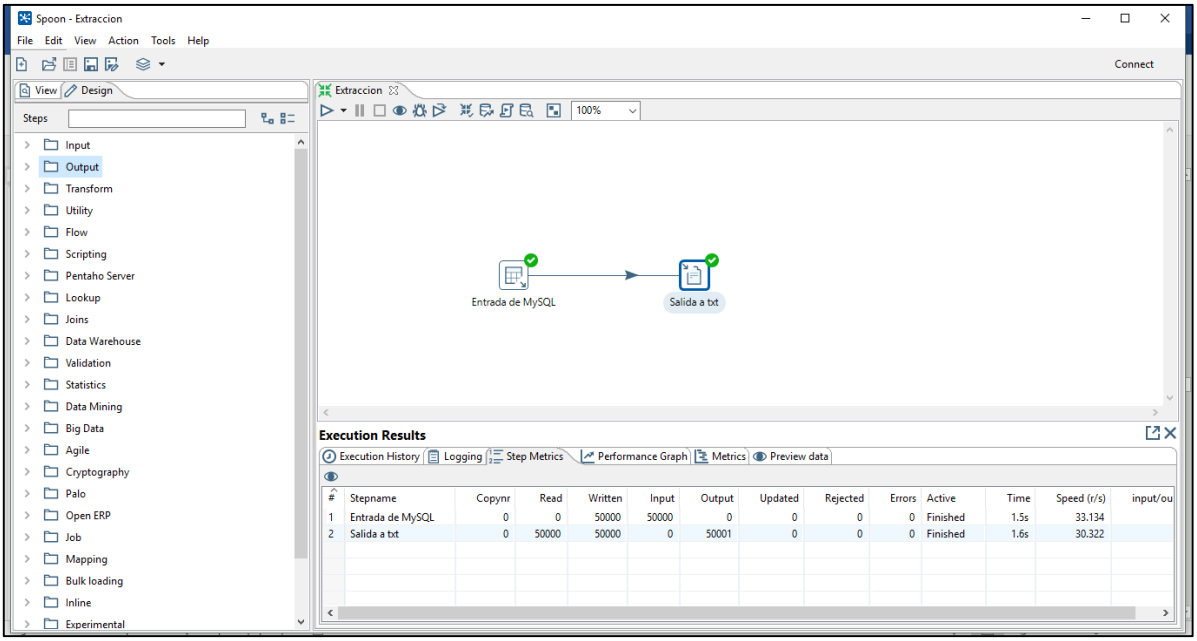


Figura 11: Extracción de datos de una tabla MySQL y Carga de datos a un archivo plano con PDI
Fuente: Elaboración Propia

En cuanto a la velocidad de lectura y escritura de datos, tiempo que tarda en ejecución del proceso y la facilidad de extracción de datos, se obtuvo que extraer datos de una tabla de MySQL y cargarlos en un archivo plano, arroja un tiempo de 1.5 segundos a una velocidad de 33.134 registros por segundo y 1.6 segundos a una velocidad de 30.322 registros por segundo respectivamente, y en cuanto a la extracción de datos se puede decir que no es proceso fácil ni sencillo en la primera interacción con la herramienta, porque se debe tener un conocimiento previo de SQL, para poder realizar la extracción de los datos, en cuanto a la conexión a la base de datos y la visualización de los datos se realiza de manera más fácil.

Se debe tener en cuenta que PDI permite guardar las conexiones a distintos Sistemas Gestores de Bases de Datos, crear tablas directamente desde el componente "**Entrada de MySQL**", por medio de SQL y por último tiene la capacidad de conectarse a 52 distintos Sistemas Gestores de Bases de Datos.

5.1.2 Extracción de datos de una tabla MySQL con TDI

Para la extracción de datos de una tabla de MySQL con TDI, primero se debe ir al contenedor "**Metadata**", y crear una conexión con MySQL en el componente "**DB Connections**", como se puede ver en la figura 12.

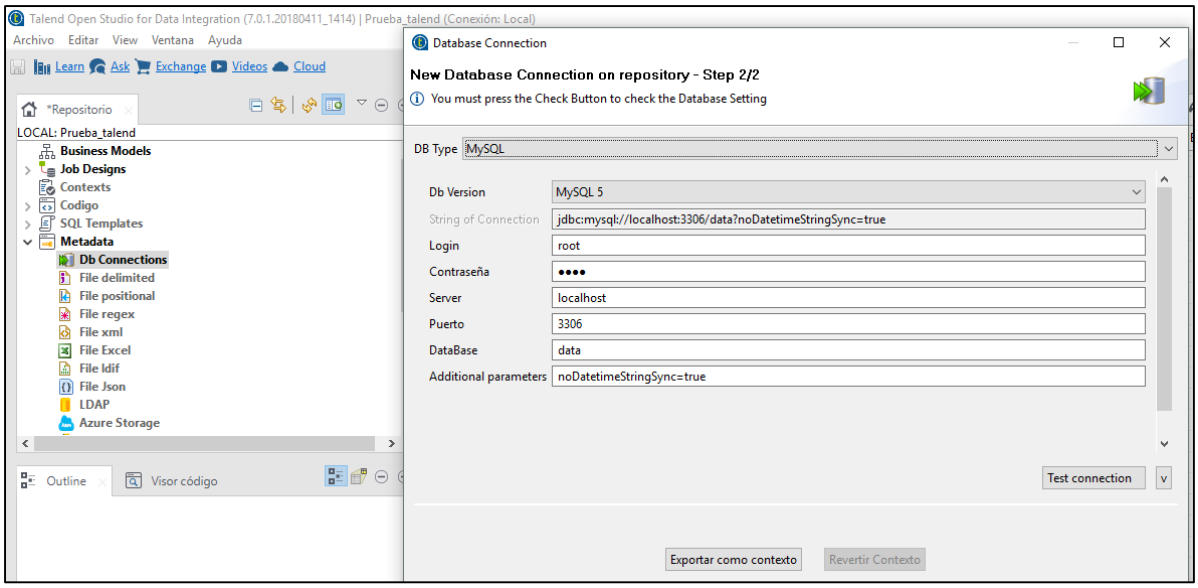


Figura 12: Conexión a la base de datos MySQL con TDI
Fuente: Elaboración Propia

Luego de haber creado la conexión con MySQL se debe ir al contenedor "**Job Designs**", y crear un "**Job**", como se ilustra en la figura 13. Que es el componente donde se realizará el proceso de extracción de datos.

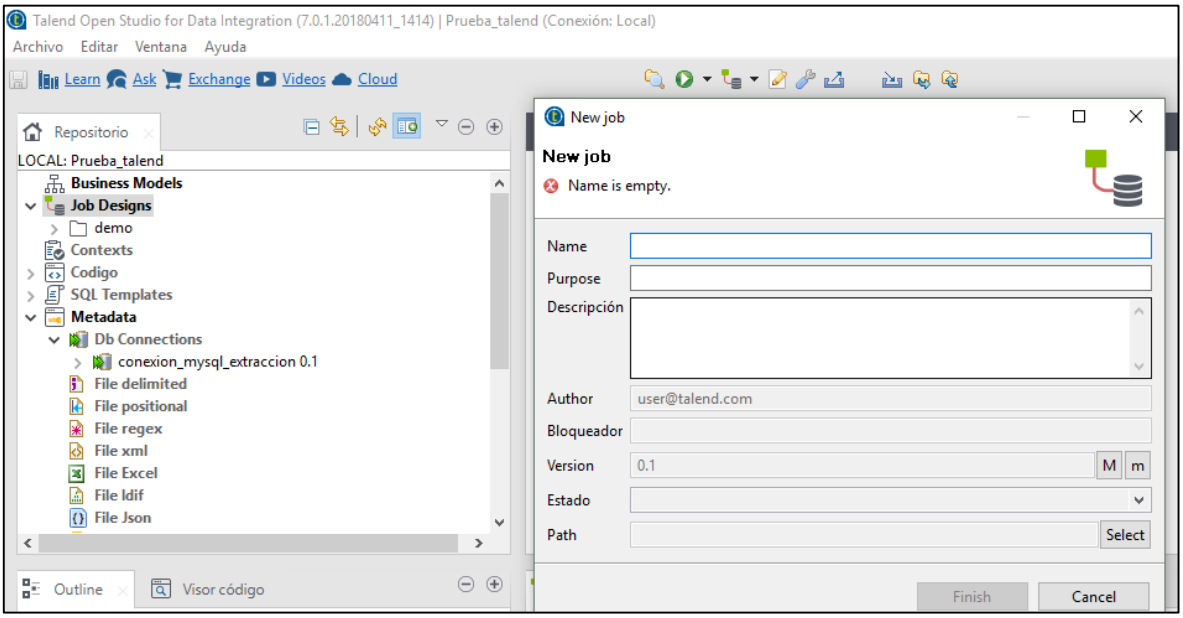


Figura 13: Creación del Job para la extracción de datos de la tabla de MySQL con TDI
Fuente: Elaboración Propia

Después de haber creado tanto la conexión como el Job, se debe importar el esquema de tablas que está contenido en la conexión "**conexión_mysql_extraccion**" en el componente "**Tables schemas**", como se puede observar en la figura 14.

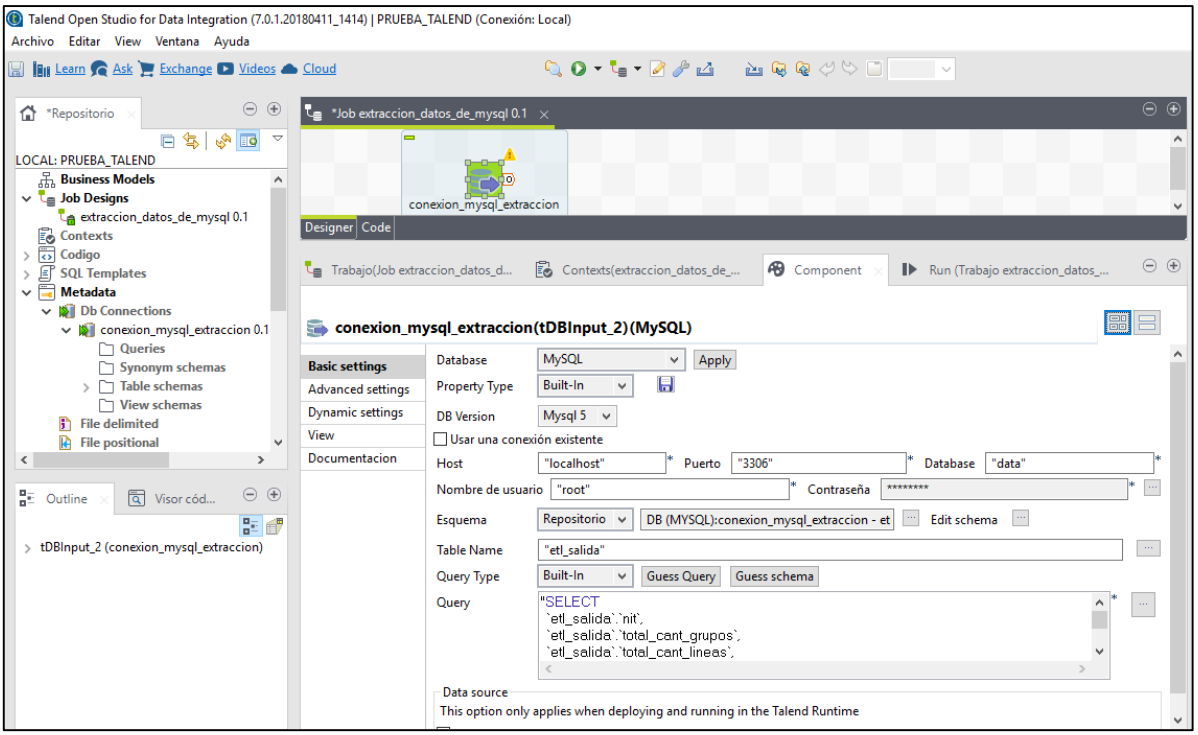


Figura 14: Importación de esquema de tablas de la conexión de MySQL con TDI
Fuente: Elaboración Propia

Continuando con el proceso se elige el componente "**tFileOutputDelimited_1**", que sirve para cargar los datos en un archivo plano, como se ve en la figura 15.

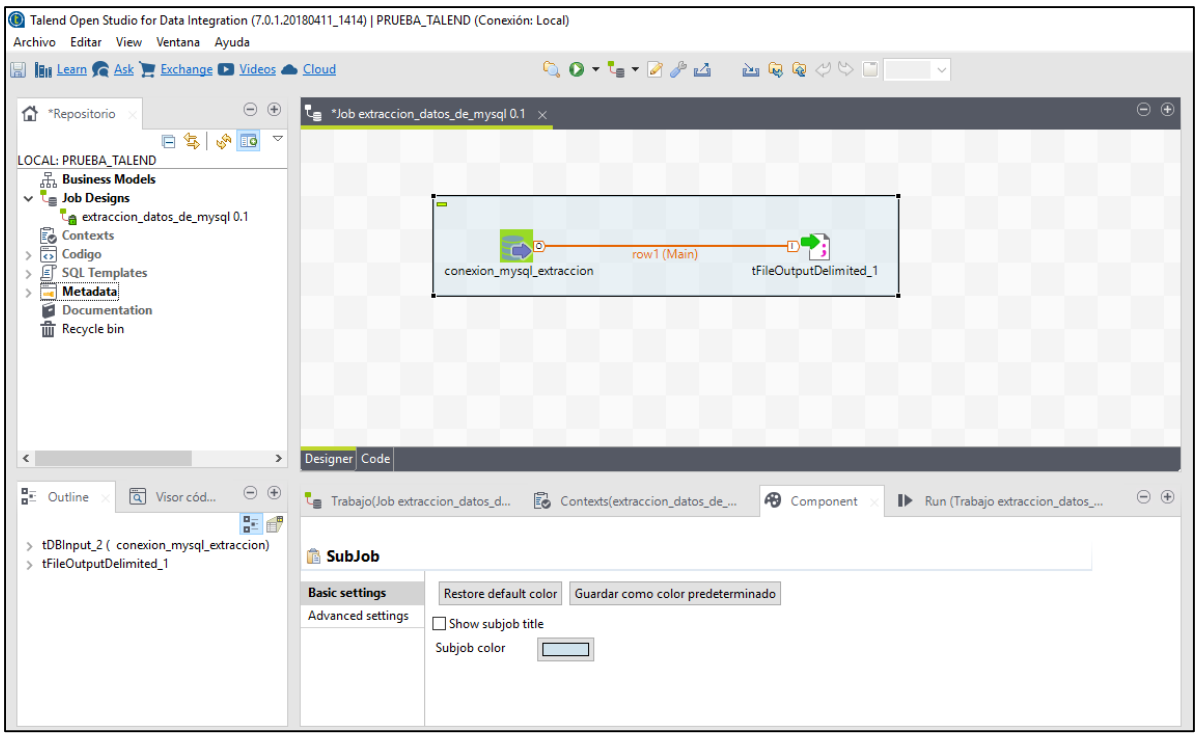


Figura 15: Componentes para la extracción de datos con TDI
Fuente: Elaboración Propia

Después se prosigue a renombrar ambos componentes "**conexión_mysql_extraccion**" como "**Entrada de MySQL**" y "**tFileOutputDelimited_1**" como "**Salida a txt**", y se ejecuta TDI, arrojando resultados, como se observa en la figura 16.

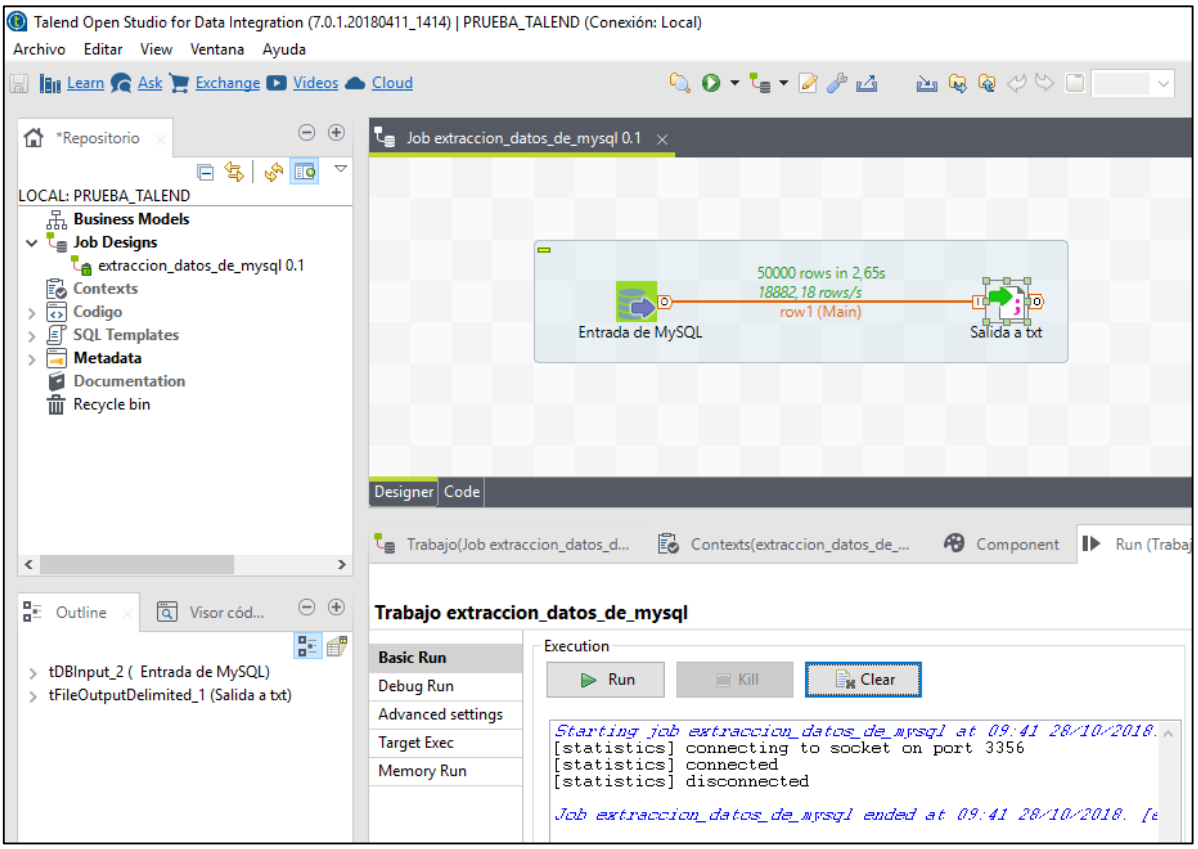


Figura 16: Extracción de datos de una tabla MySQL y Carga de datos a un archivo plano con TDI
Fuente: Elaboración Propia

La figura anterior, presenta como resultado que al extraer los datos de una tabla de MySQL y cargarlos en un archivo plano, tardó un tiempo de 2.65 segundos con una velocidad de 18.882 registros por segundo, en el proceso, y en cuanto a la extracción de los datos se puede decir que no es un proceso sencillo, porque primero hay que generar un componente en el contenedor "**Metadata**" que contiene la información de la conexión, que luego se utiliza para importar el esquema de las tablas, y que por último se utiliza para la extracción de los datos en el componente "**Job**", por medio de SQL.

También se debe tener en cuenta que TDI permite guardar las conexiones a distintos Sistemas Gestores de Bases de Datos, crear tablas directamente desde el componente "**Entrada de MySQL**" y permite establecer conexión a 39 distintos Sistemas Gestores de Bases de Datos.

5.1.3 Extracción de datos de una tabla MySQL con OR

Para la extracción de datos de una tabla de MySQL con OR, primero se debe ir a la opción "**Database**", donde luego se elige el tipo de base de datos a utilizar, en este caso MySQL, como se observa en la figura 17.

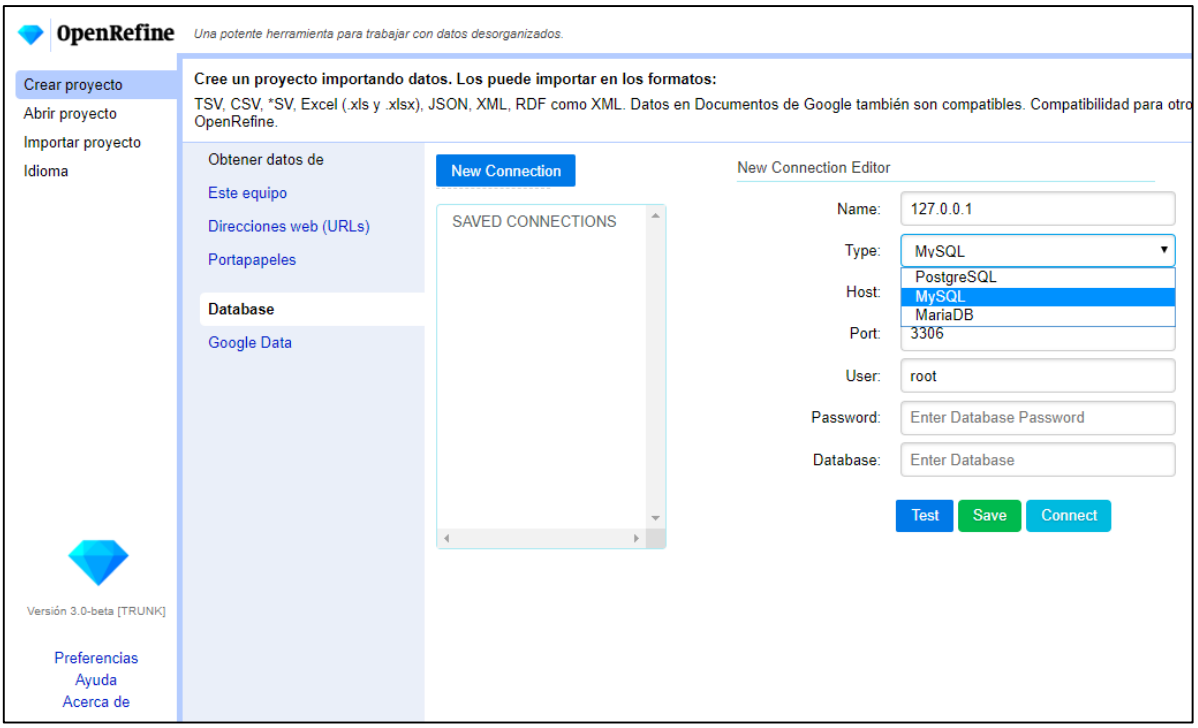


Figura 17: Conexión a la base de datos MySQL con OR
Fuente: Elaboración Propia

Después se prosigue a crear la conexión con la base de datos y guardarla, como se ilustra en la figura 18.

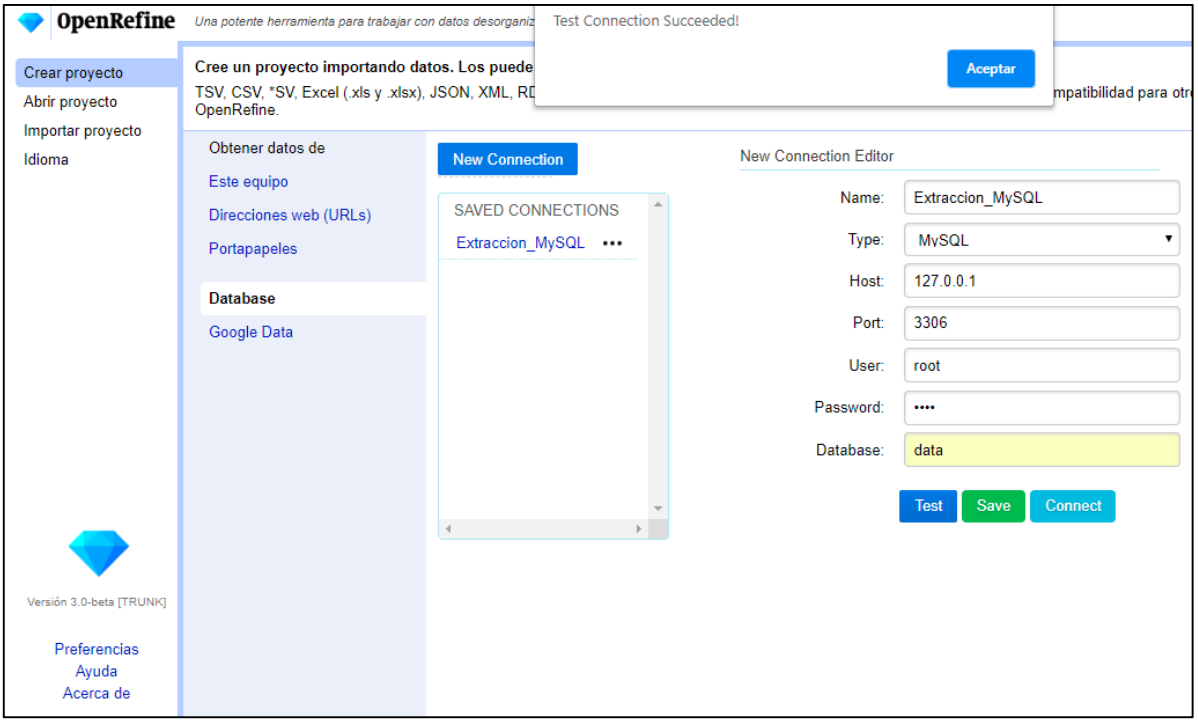


Figura 18: Conexión a MySQL creada y guardada con OR
Fuente: Elaboración Propia

Luego de crear y guardar la conexión se prosigue a realizar la consulta SQL, como se puede ver en la figura 19, que es la consulta estructurada para extraer los datos.

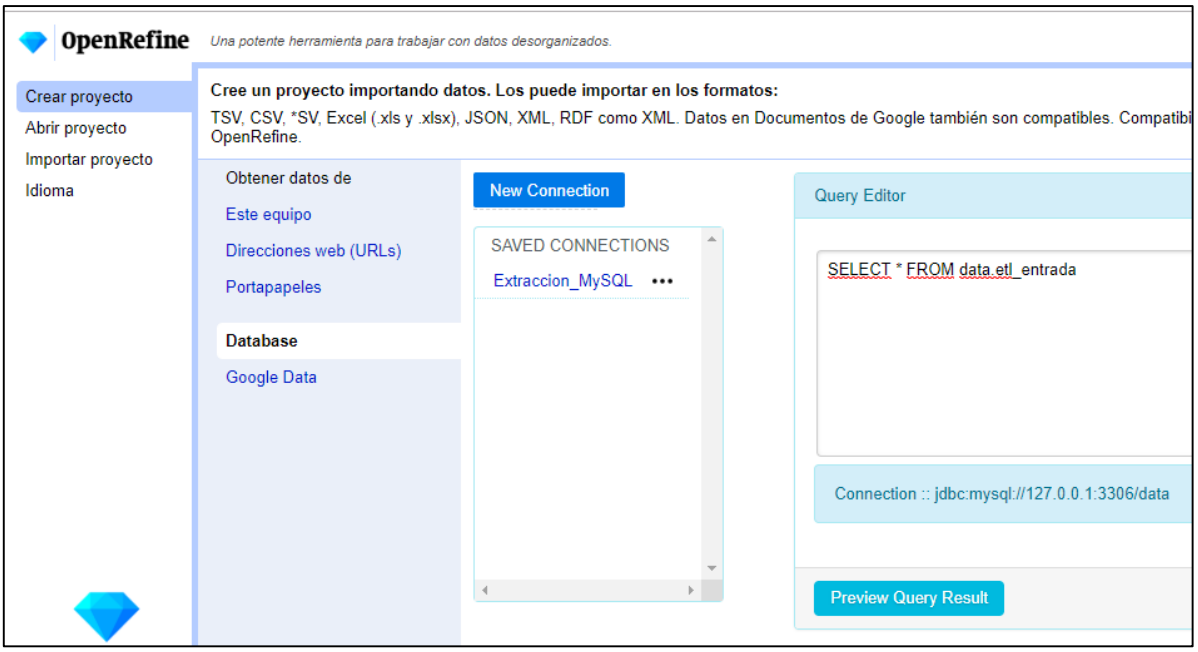


Figura 19: Consulta SQL para la extracción de los datos de MySQL con OR
Fuente: Elaboración Propia

La figura 20, contiene los datos extraídos al realizar la consulta SQL a la tabla de base de datos MySQL.

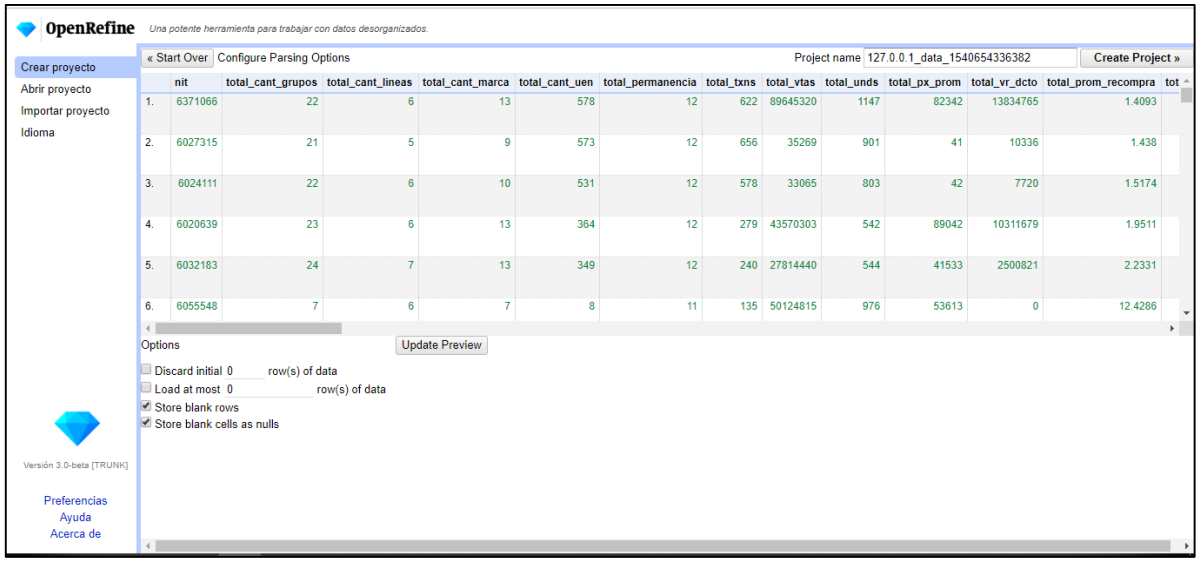


Figura 20: Datos extraídos por OR de MySQL antes de crear el proyecto
Fuente: Elaboración Propia

Una vez extraídos los datos, se procede a crear un proyecto con el botón "**Create Project**", donde primero se le debe indicar un nombre del proyecto. El último paso es exportar los datos extraídos al formato deseado como se ilustra en la figura 21.

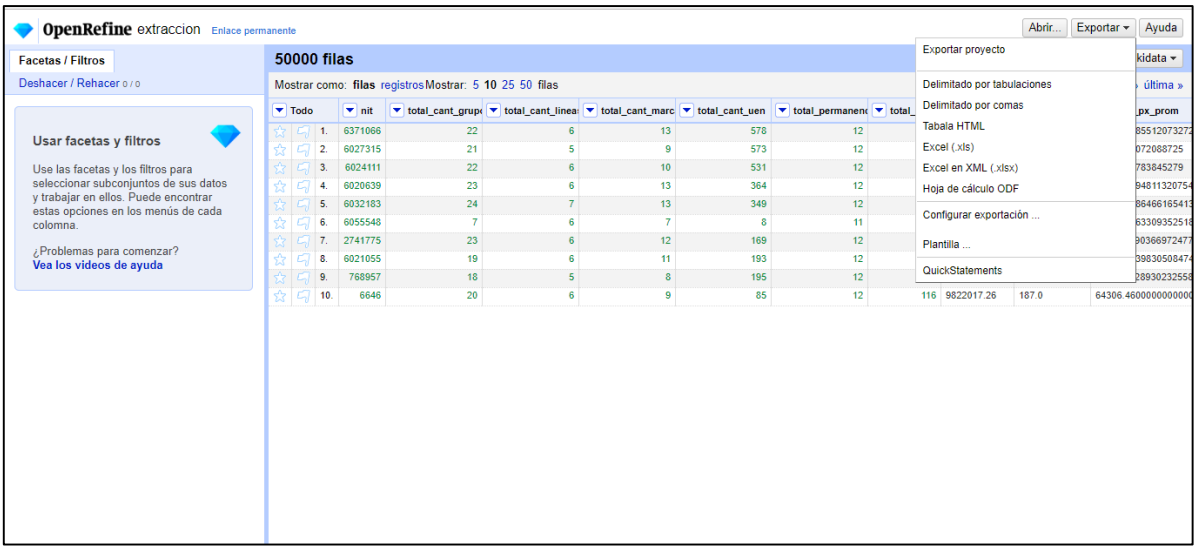


Figura 21: Datos extraídos por OR de MySQL después de crear el proyecto
Fuente: Elaboración Propia

Para el caso de OR, en cuanto a la velocidad de lectura y escritura de datos, tiempo que tarda en ejecución del proceso y facilidad de extracción de datos, no es posible medir de una forma fácil como en PDI o TDI, porque no posee la capacidad de llevar el registro histórico de lectura y escritura de datos con su respectivo tiempo de ejecución, pero la conexión a base de datos es muy sencilla con respecto a PDI y TDI, dándole una ventaja en cuanto a la forma de conectarse a la base de datos MySQL. La extracción de los datos se realiza de forma rápida por medio de SQL, extrayendo los datos requeridos para el procesamiento. Una de las ventajas que posee OR, es que se pueden observar los datos en forma de tabla como se observa en la figura 21.

Se debe tener en cuenta que OR, permite guardar las conexiones a distintos Sistemas Gestores de Bases de Datos, y que solo permite establecer conexión con 3 distintos Sistemas Gestores de Bases de Datos, en este caso no te permite crear una tabla desde SQL.

5.1.4 Conclusión de la extracción de datos de una tabla de MySQL con PDI, TDI y OR

En cuanto a la velocidad de lectura y escritura de datos, tiempo que tarda en ejecución del proceso y facilidad de extracción de datos de una tabla MySQL, PDI fue la más eficiente en velocidad y tiempo de ejecución, mientras que OR cuenta con la facilidad en la extracción de datos, por la forma sencilla e intuitiva que permite extraer datos sin gran esfuerzo ni conocimientos previos. Tanto PDI, TDI y OR tienen en común que pueden extraer datos con sentencias SQL. Para el caso de TDI se debe crear un esquema de metadatos primero antes de poder procesar los datos, algo que no sucede con PDI y OR. También cuentan con la facilidad de poder visualizar los datos, es decir en OR se visualizan los datos como si fuera tipo Excel, mientras que PDI los visualizan fácil solo dando clic en el botón “**Preview**” de cada componente de entrada de datos, mientras que en TDI se visualizan utilizando un componente llamado “**tLogRow**”.

5.2 Extracción de datos de un archivo plano

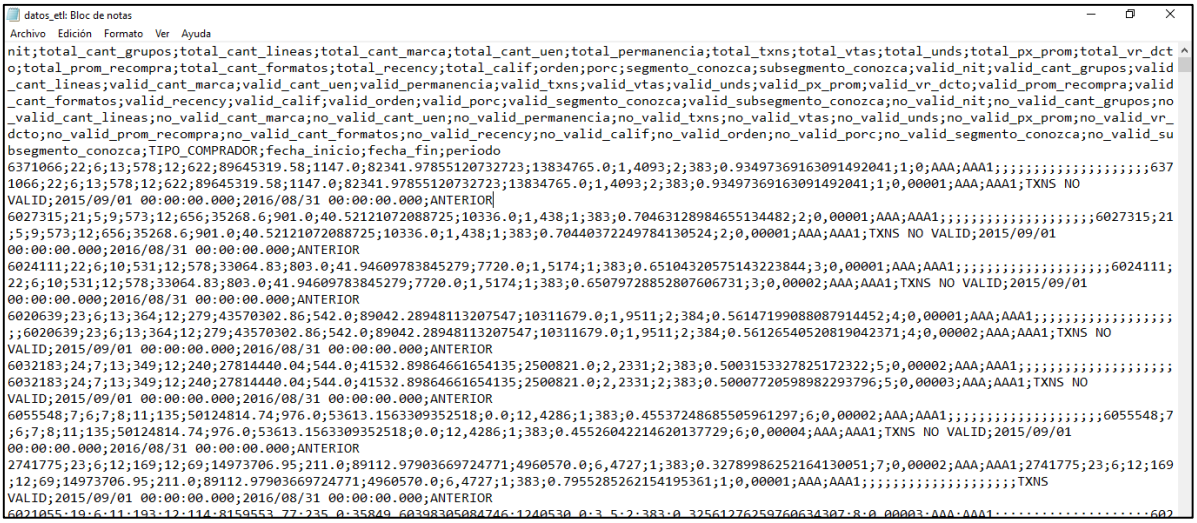


Figura 22: Datos del archivo plano
Fuente: Elaboración Propia

Para comparación de PDI, TDI y OR, en cuando a la velocidad de lectura y escritura de datos, tiempo que tarda en ejecución del proceso y facilidad de extracción de datos, se usará un archivo plano con 50.000 registros y 61 campos, que se ven en la figura 22. A continuación se describe como extraer datos de un archivo plano con PDI, TDI y OR.

5.2.1 Extracción de datos de un archivo plano con PDI

Para la extracción de datos de un archivo plano con PDI, primero se debe crear una transformación. Se debe ir a "**File**", luego a "**New**", y dar clic en la opción "**Transformation**", donde luego se va al contenedor "**Input**", que contiene los componentes de entrada de datos en PDI, y se arrastra el componente "**Text file input**", que sirve para establecer una conexión con el archivo plano, como se puede ver en la figura 23.

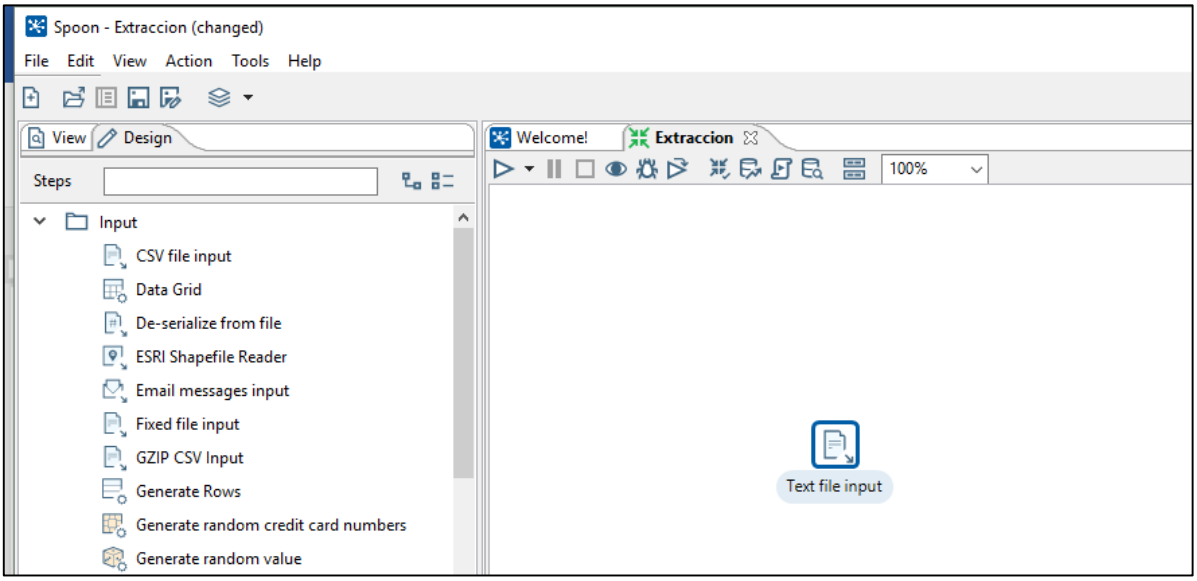


Figura 23: Componente de entrada de datos del archivo plano con PDI
Fuente: Elaboración Propia

Al haber establecido la conexión con el archivo plano, se prosigue a indicar el componente de salida de datos, se va al contenedor "**Output**", y se arrastra el componente "**Microsoft Excel Output**", que se renombra como "**Salida a excel de txt 2**", que sirve para cargar los datos en un archivo de Excel, como se ilustra en la figura 24, donde se observa el resultado de la extracción de datos del archivo plano.

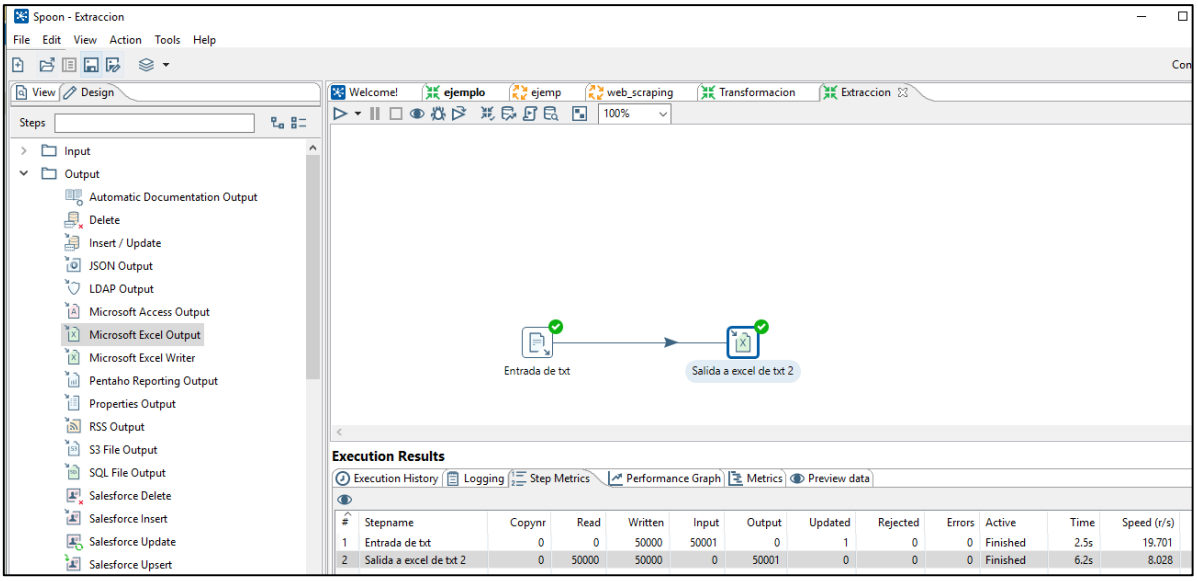


Figura 24: Extracción de datos de un archivo plano y Carga de datos a un archivo de Excel con PDI
Fuente: Elaboración Propia

En cuanto a la velocidad de lectura y escritura de datos, tiempo que tarda en ejecución del proceso y facilidad de extracción de datos, como resultado se observa que PDI extrajo los datos, y los cargo en un archivo Excel, en un tiempo de 2.5 segundos a una velocidad de 19.701 registros por segundo y 6.2 segundos a una velocidad de 8.028 registros por segundo respectivamente, y en cuanto a la facilidad de extraer datos, fue sencillo y fácil porque permite leer el directorio donde está ubicado el archivo plano que contiene los datos a extraer y permite cargar los datos de una forma relativamente fácil a un archivo de Excel.

5.2.2 Extracción de datos de un archivo plano con TDI

Para la extracción de datos de un archivo plano con TDI, primero se debe ir al contenedor "**Metadata**", y crear una conexión con el archivo plano en el componente "**File delimited**", como se observa en la figura 25.

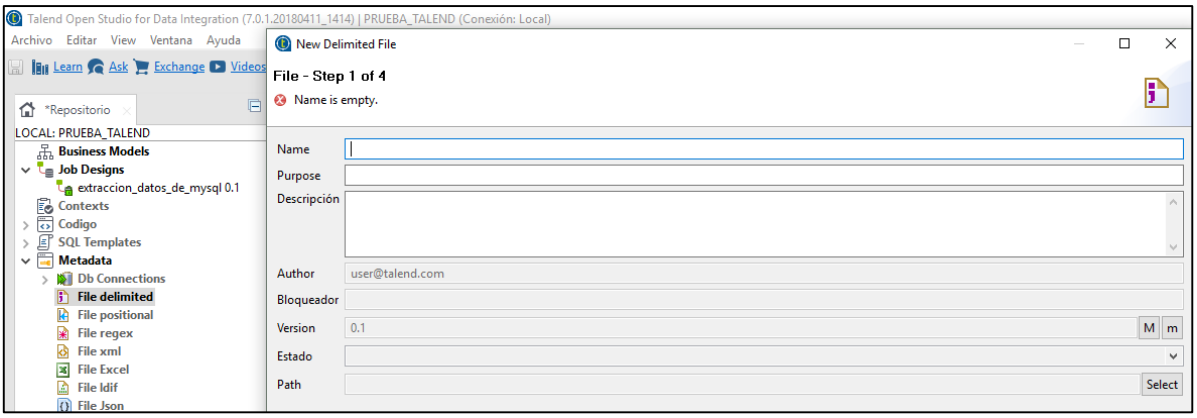


Figura 25: Conexión al archivo plano con TDI
Fuente: Elaboración Propia

Luego de haber creado la conexión con el archivo plano se debe ir al contenedor "**Job Designs**", y crear un "**Job**", como se puede ver en la figura 26, que sirve para realizar el proceso de extracción de datos.

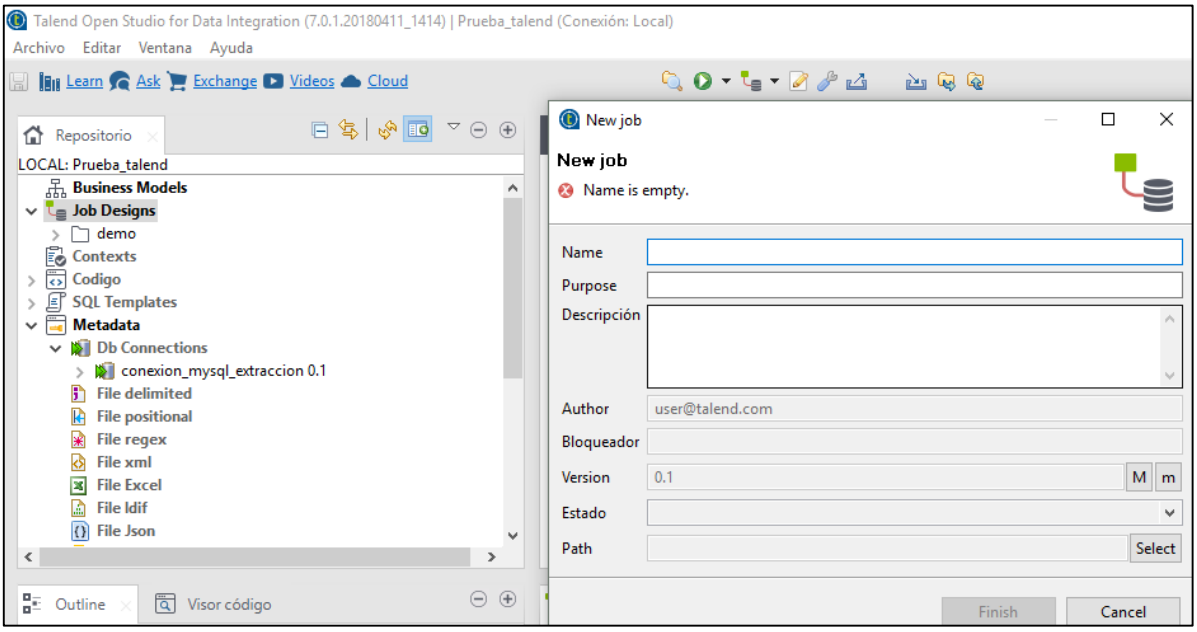


Figura 26: Creación del Job de extracción de datos del archivo plano con TDI
Fuente: Elaboración Propia

Después de haber creado tanto la conexión como el Job, se debe indicar a dónde se van a cargar los datos, en este caso a un archivo de Excel en el componente llamado "**Salida a Excel de txt**", por último, se inicia el proceso de extracción de datos del archivo plano, como se ilustra en la figura 27.

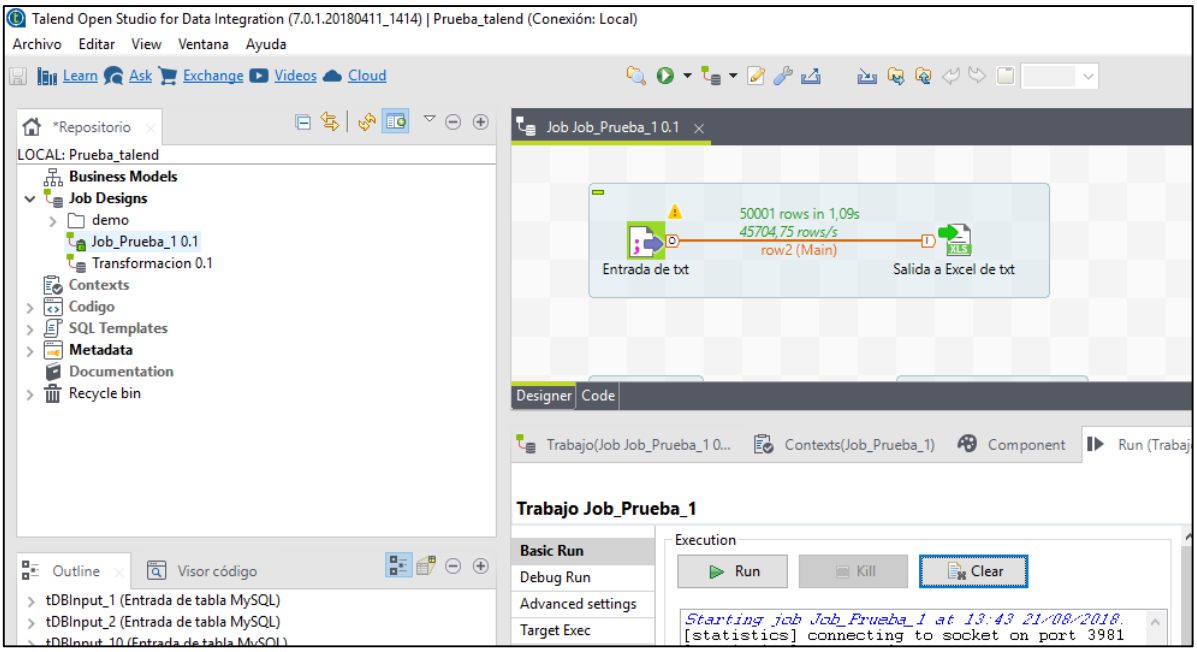


Figura 27: Extracción de datos de un archivo plano y Carga de datos a un archivo de Excel con TDI
Fuente: Elaboración Propia

En cuanto a la velocidad de lectura y escritura de datos, tiempo que tarda en ejecución del proceso y facilidad de extracción de datos, arroja como resultado, que al realizar el proceso de extracción de datos de un archivo plano y cargarlos en un archivo Excel, tardo 1.09 segundos a una velocidad de 45.704 registros por segundo en todo el proceso, y en cuanto a la facilidad de extraer datos, no fue tan sencillo, porque se debe crear un esquema de metadatos antes de poder extraer los datos del directorio donde está ubicado el archivo plano, que luego se configura para darle el formato de cada campos.

5.2.3 Extracción de datos de un archivo plano con OR

Para la extracción de datos de un archivo plano con OR, primero se debe ir a la opción "**Este equipo**", donde luego se busca en el directorio el archivo plano, como se observa en la figura 28.

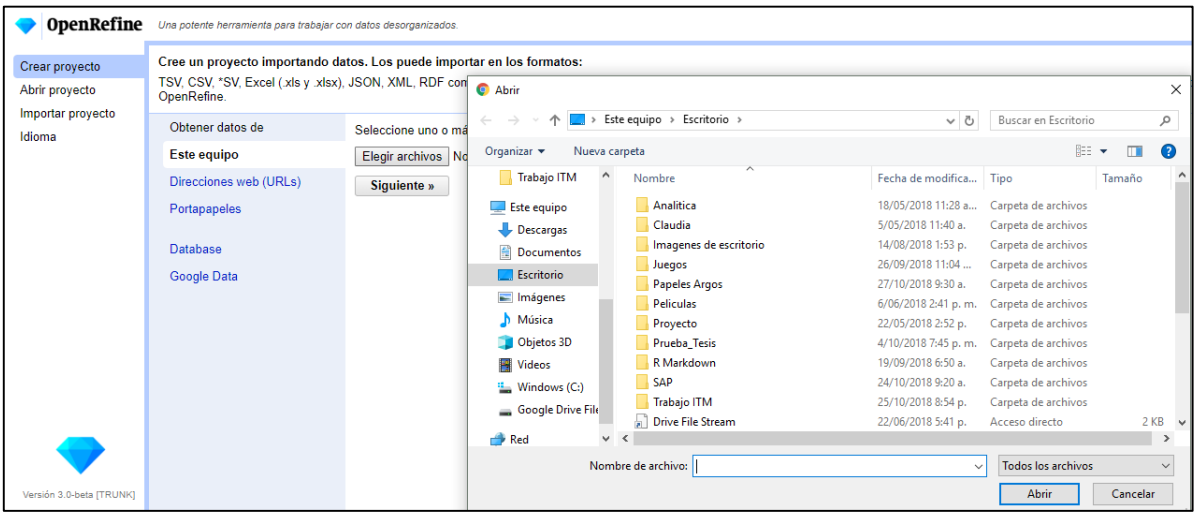


Figura 28: Conexión al archivo plano con OR
Fuente: Elaboración Propia

Luego de extraer los datos del archivo plano, se prosigue a crear el proyecto con el nombre de “extracción”, como se puede ver en la figura 29.

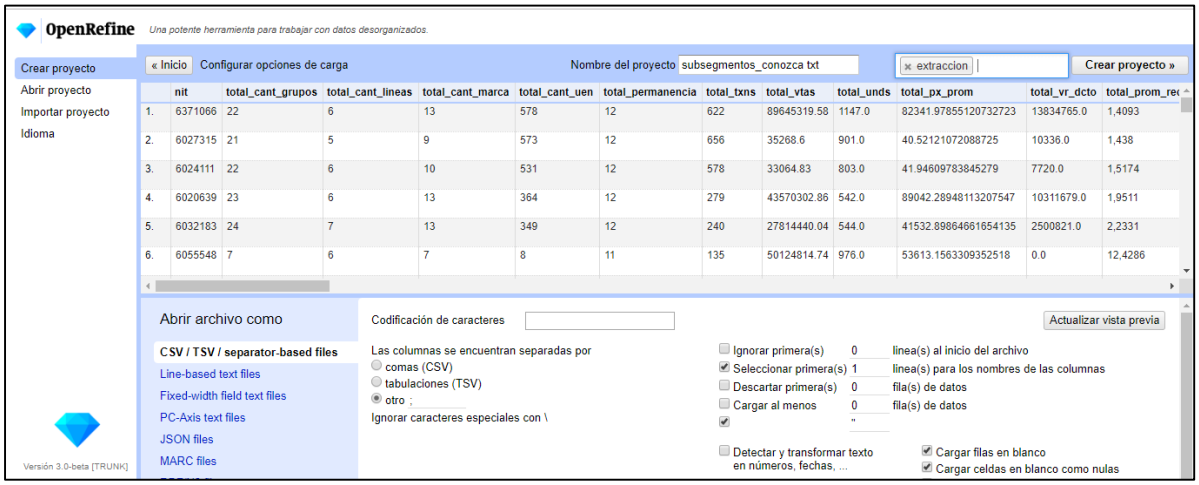


Figura 29: Creación del proyecto para la extracción de datos de un archivo plano con OR
Fuente: Elaboración Propia

Por último, en OR, al realizar el proceso de extracción de datos de un archivo plano, para cargarlos a un archivo Excel, no permite llevar el registro del tiempo ni la velocidad de lectura y escritura de datos, como en las demás herramientas al realizar la extracción, pero a la hora de cargar los datos a un archivo de Excel, el servidor se cayó varias veces, es decir no soporto la carga de 50.000 registros y 61 campos, como se puede ver en la figura 30 y 31. Es decir OR cuenta con una desventaja, y es que solo procesa datos de tamaños pequeños, para el caso fueron (50.000 registros y 61 campos), ya que se consume todos los recursos del sistema como CPU y RAM, provocando que el rendimiento sea bajo, llevando a que el proceso tarde más de los normal, es decir si desea extraer o cargar datos con OR, lo esencial es que cuente con computador que posea buen procesador mínimo Core i7 y RAM mínimo 8gb.

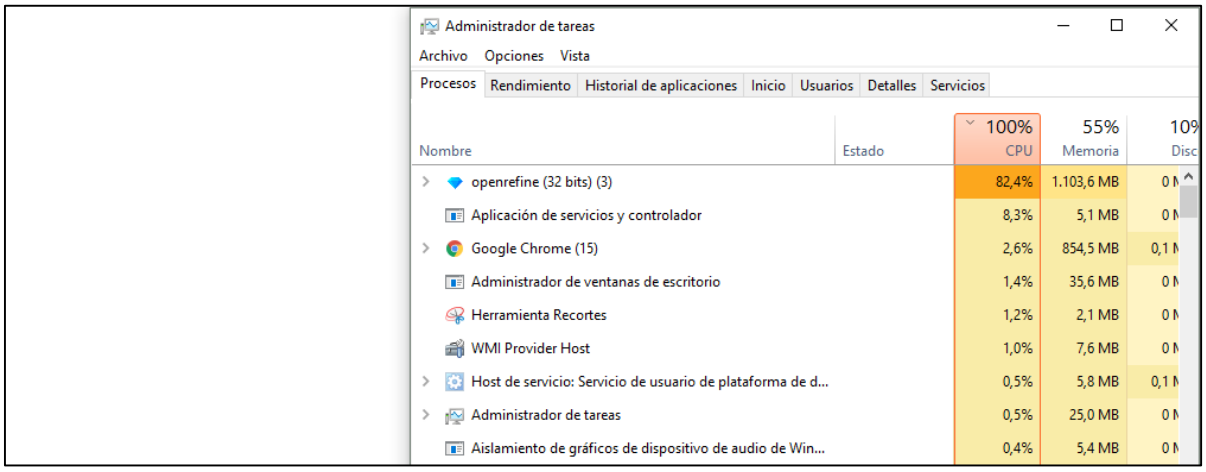


Figura 30: Consumo de CPU y RAM en la carga de datos a un archivo de Excel con OR
Fuente: Elaboración Propia

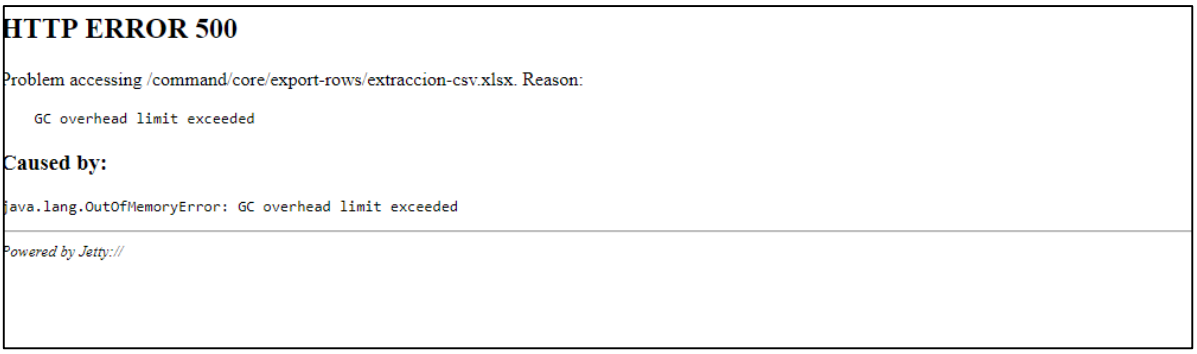


Figura 31: Una imagen del sistema caído en la carga de datos a un archivo de Excel con OR
Fuente: Elaboración Propia

5.2.4 Conclusión de la extracción de datos de un archivo plano con PDI, TDI y OR

En cuanto a la velocidad de lectura y escritura de datos, tiempo que tarda en ejecución del proceso y facilidad de extracción de datos, TDI fue el más eficiente en velocidad y tiempo de ejecución, mientras que OR cuenta con la facilidad en la extracción de datos, porque solo debes buscar el directorio del archivo plano e indicar el separador de campos. PDI, TDI y OR tienen en común que pueden extraer datos de archivos planos.

5.3 Conclusión de extracción de datos con PDI, TDI y OR

OR es la herramienta que posee una gran facilidad para extraer datos, mientras que PDI posee un mejor rendimiento en cuanto a la velocidad y tiempo de ejecución a la hora de extraer datos de una tabla de MySQL y TDI posee un mejor rendimiento en cuanto a velocidad y tiempo de ejecución a la hora de extraer datos de un archivo plano. En la tabla 5.1 se puede observar una comparación de las herramientas en proceso de extracción usando estos datos particulares, que permite identificar ventajas y desventajas.

Herramienta / Proceso	Talend Data Integration (TDI)	Pentaho Data Integration (PDI)	OpenRefine (OR)
Cantidad de Sistemas Gestores de Bases de Datos a los que se puede conectar	39	52	3
Facilidad de extracción de datos de una tabla de MySQL	Facilidad baja	Facilidad media	Facilidad alta
Tiempo de extracción de datos de una tabla de base de datos MySQL que se envían a un archivo plano (txt) (50.000 registros y 61 campos)	Rápido (2.65 Segundos)	Rápido (1.5 Segundos)	No es medible dentro del programa como lo hacen las demás herramientas
Velocidad de extracción de datos de una tabla de base de datos MySQL que se envían a un archivo plano (txt) (50.000 registros y 61 campos)	Lee datos a una velocidad de 18.882 registros por segundo	Lee datos a una velocidad de 33.134 registros por segundo	No es medible dentro del programa como lo hacen las demás herramientas

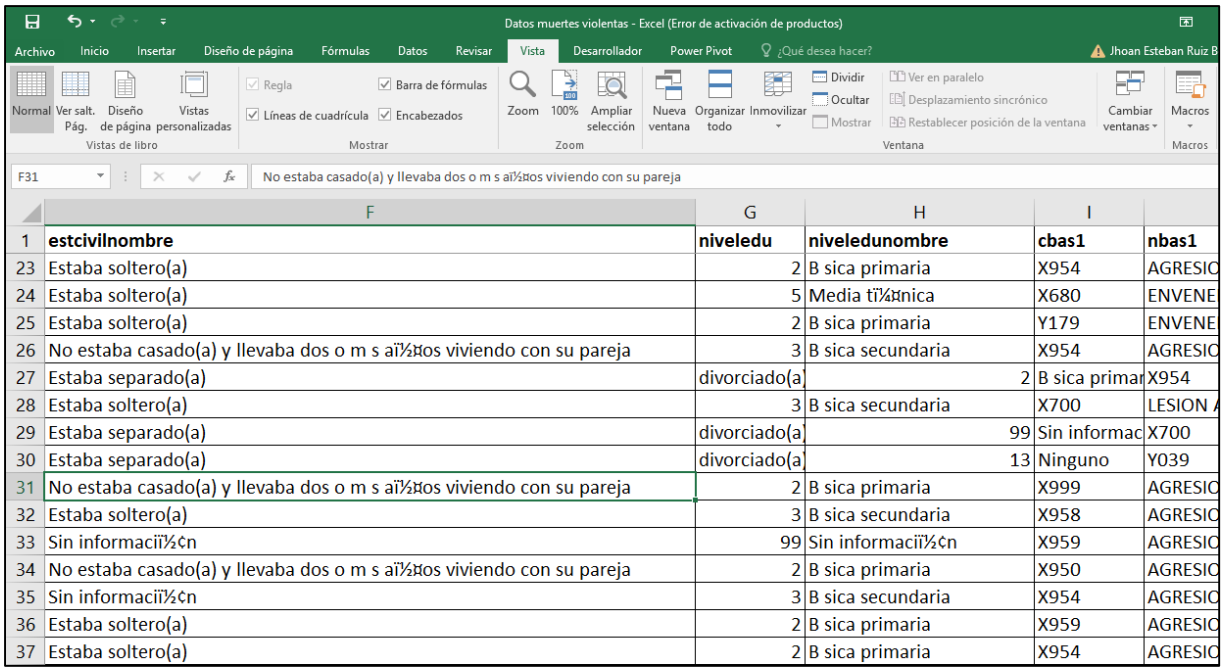
Facilidad de extracción de datos de un archivo plano	Facilidad baja	Facilidad media	Facilidad alta
Tiempo de extracción de datos de un archivo plano (txt) que se envían a un archivo de Excel (50.000 registros y 61 campos)	Rápido (1.09 Segundos)	Rápido (2.5 Segundos)	No es medible dentro del programa como lo hacen las demás herramientas
Velocidad de extracción de datos de un archivo plano (txt) que se envían a un archivo de Excel (50.000 registros y 61 campos)	Lee datos a una velocidad de 45.704 registros por segundo	Lee datos a una velocidad de 19.701 registros por segundo	No es medible dentro del programa como lo hacen las demás herramientas
Necesita generar un esquema de metadatos para la extracción de datos	Si (Metadatos)	No	No
Google Analytics	No	Si	No
Salesforce	Si	Si	No
Email	Si	Si	No
Archivo TVS	No	No	Si
Archivo RDF	No	No	Si
Archivo ARFF	Si	No	No
SAS	No	Si	No
SAP	No	Si	No
Nube	**Extracción de datos de la Nube**		
Azure Storage	Si	No	No
Google Storage	Si	No	No
Amazon	Si	No	No
Big Data	**Extracción de datos de Big Data**		
Hive	Si	No	No
BigQuery	Si	No	No
Avro	No	Si	No
Cassandra	No	Si	No
CouchDb	No	Si	No
MongoDB	No	Si	No
HBase	No	Si	No
Hadoop	No	Si	No

Tabla 5-1: Comparación de las herramientas en la componente de Extracción de datos
Fuente: Elaboración Propia

En la tabla 5.1, se puede notar que PDI y TDI, no solo extraen datos de una base de datos o un archivo plano, sino que también pueden extraer datos de muchas fuentes, como bases de datos no relacionales, archivos XML, archivos de Excel, JSON, la nube entre otros. Mientras que OR es más limitado a la hora de extraer datos ya que solo puede extraer datos de algunas bases de datos, y archivos planos.

6 Comparación de herramientas en el proceso de transformación de datos (T)

Para el ejemplo con PDI, TDI y OR, que tienen muchos procesos de transformación, el ejemplo práctico se realizará transformando un archivo de Excel llamado “Datos Muertes Violentas” que fueron descargados de la Web [51]. Se muestran algunos datos en la figura 32.



	F	G	H	I	
1	**estcivilnombre**	**niveledu**	**niveledunombre**	**cbas1**	**nbas1**
23	Estaba soltero(a)	2	B sica primaria	X954	AGRESIO
24	Estaba soltero(a)	5	Media tï¼¤nica	X680	ENVENE
25	Estaba soltero(a)	2	B sica primaria	Y179	ENVENE
26	No estaba casado(a) y llevaba dos o m s aï½¤os viviendo con su pareja	3	B sica secundaria	X954	AGRESIO
27	Estaba separado(a)	divorciado(a	2	B sica primar	X954
28	Estaba soltero(a)	3	B sica secundaria	X700	LESION A
29	Estaba separado(a)	divorciado(a	99	Sin informac	X700
30	Estaba separado(a)	divorciado(a	13	Ninguno	Y039
31	No estaba casado(a) y llevaba dos o m s aï½¤os viviendo con su pareja	2	B sica primaria	X999	AGRESIO
32	Estaba soltero(a)	3	B sica secundaria	X958	AGRESIO
33	Sin informaciï½¢n	99	Sin informaciï½¢n	X959	AGRESIO
34	No estaba casado(a) y llevaba dos o m s aï½¤os viviendo con su pareja	2	B sica primaria	X950	AGRESIO
35	Sin informaciï½¢n	3	B sica secundaria	X954	AGRESIO
36	Estaba soltero(a)	2	B sica primaria	X959	AGRESIO
37	Estaba soltero(a)	2	B sica primaria	X954	AGRESIO

Figura 32: Archivo de Excel para el proceso de transformación
Fuente: Elaboración Propia

En la figura 32, se puede observar que tiene problemas de caracteres especiales como "**Sin informaciï½¢n**" y "**Media tï¼¤nica**", entre otros. También tiene problemas con registros que están en columnas a las que no corresponden, como "**divorciado (a)**", que está en el campo "**niveledu**", el cual debería estar en el campo "**estcivilnombre**", los cuales se deberán filtrar, y luego concatenar para poder unir dos columnas, y ubicar los datos de las columnas concatenadas en su campo correspondiente.

Como se busca comparar a PDI, TDI y OR, en el proceso de transformación de datos, en cuando a la velocidad de transformación, tiempo que tarda en ejecución del proceso y la facilidad de la transformación de datos, utilizando un archivo de "Datos Muertes Violentas", con 41.867 registros y 10 campos, para realizar las transformaciones pertinentes con el propósito de dejar el archivo limpio para su posterior carga en un archivo de Excel.

6.1 Transformación de los datos de muertes violentas con PDI

Para realizar el proceso de transformación en PDI de los datos de muertes violentas, primero se debe crear una transformación. Para esto se debe ir a "**File**", luego a "**New**", y dar clic en la opción "**Transformation**", como se ve en la figura 33.

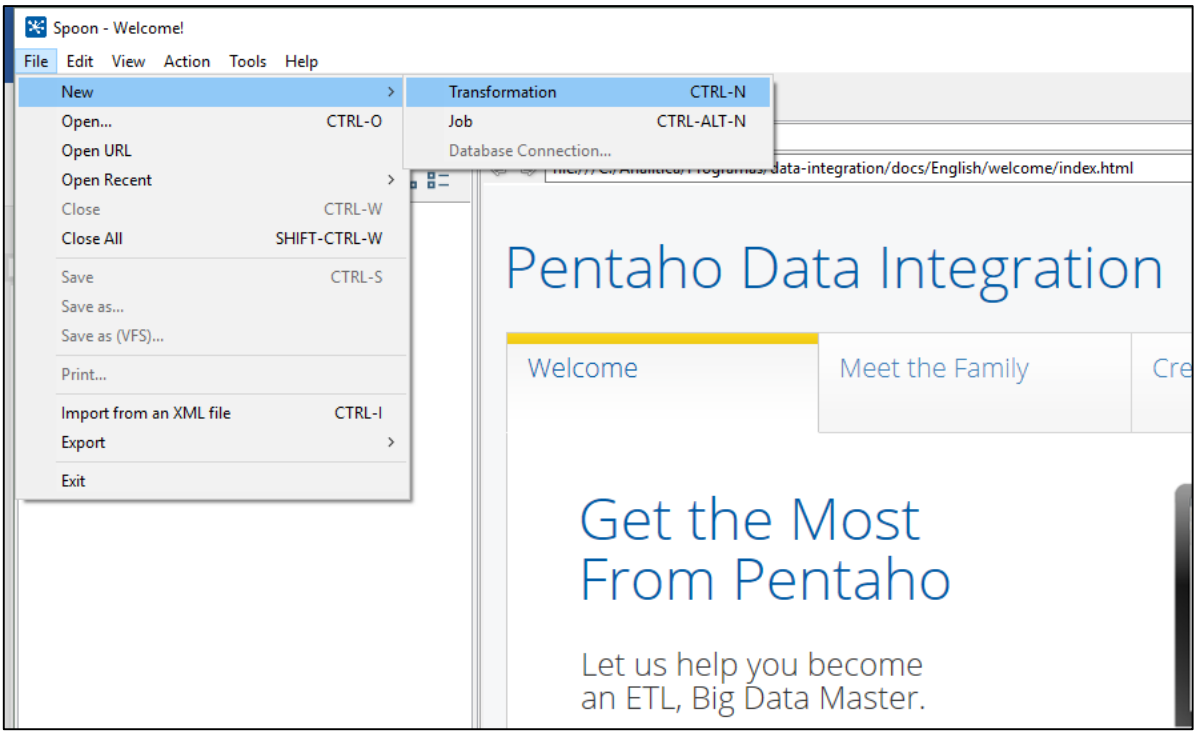


Figura 33: Creación de una transformación en PDI
Fuente: Elaboración Propia

Luego de crear la transformación, se debe ir al contenedor "**Input**" y arrastrar el componente "**Microsoft Excel input**", que se renombra como "**Entrada Muertes Violentas**", que sirve para extraer los datos que están en un archivo de Excel, como se ilustra en a figura 34.

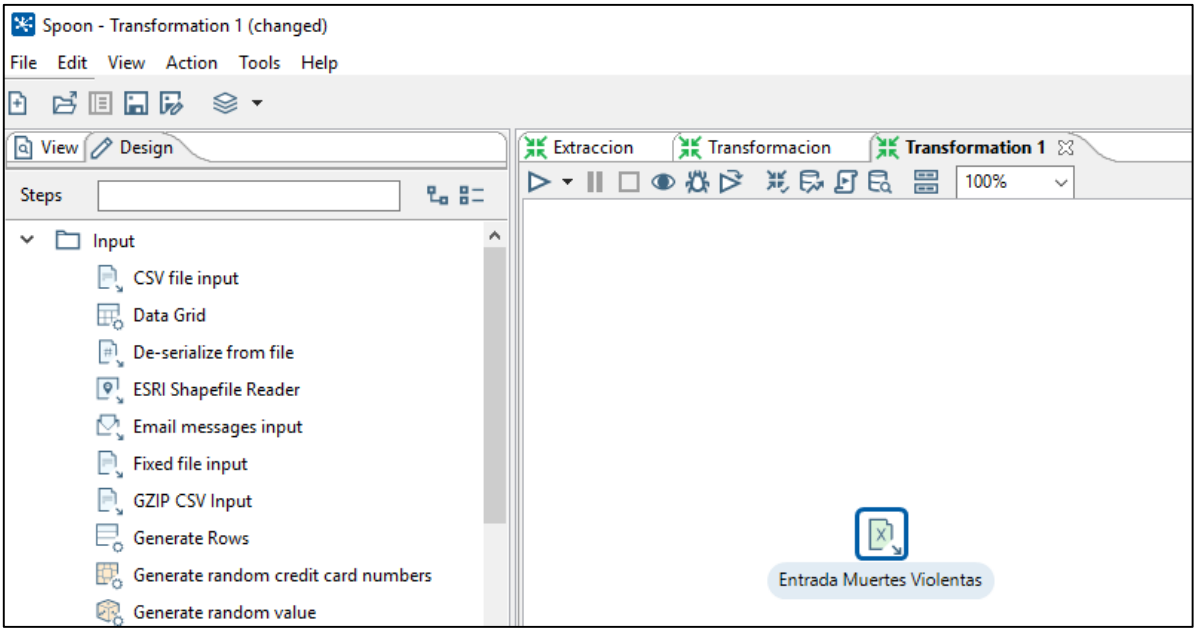


Figura 34: Extracción de datos de muertes violentas con PDI
Fuente: Elaboración Propia

Después de extraer los datos, se prosigue a realizar las transformaciones correspondientes. Primero se filtran los registros que están en campos que no corresponden a su tipo de datos, se debe ir al contenedor "**Flow**", que contiene los flujos de datos, y se arrastra el componente "**Filter rows**", que se renombra como "**Filtrar Filas**", como se puede ver en la figura 35.

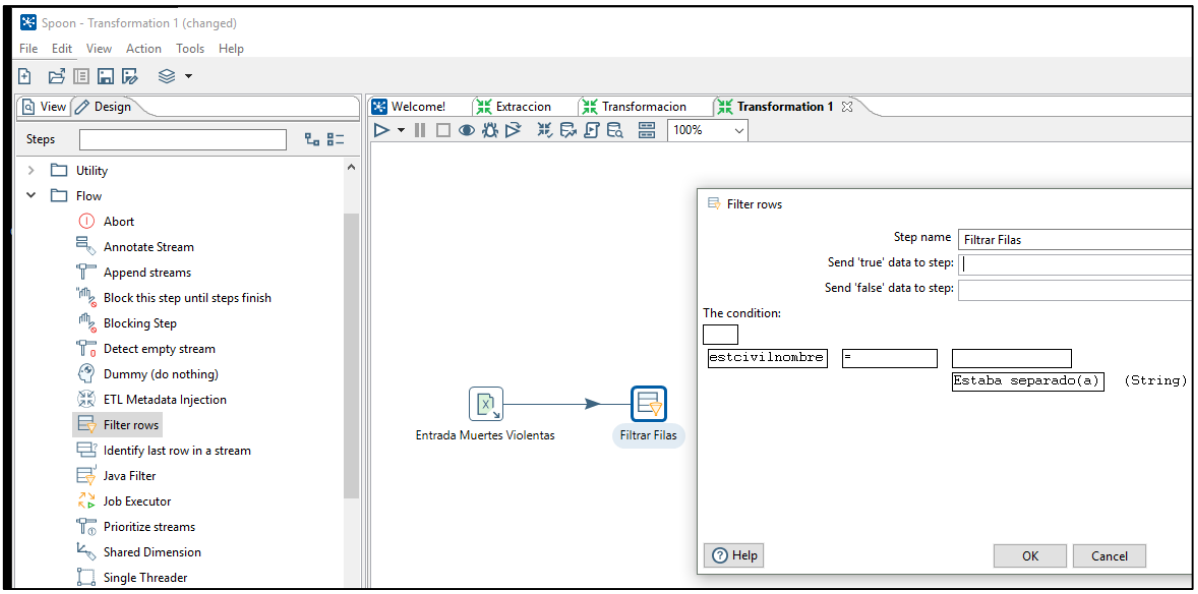


Figura 35: Componente para filtrar las filas con PDI
Fuente: Elaboración Propia

A continuación, se reemplaza los caracteres especiales que poseen algunos registros, se va al contenedor "**Transform**", que contiene las transformaciones, y se arrastra el componente "**Replace in string**", que se renombra como "**Reemplaza caracteres especiales**" y el otro como "**Reemplaza caracteres especiales 1**", como se observa en la figura 36.

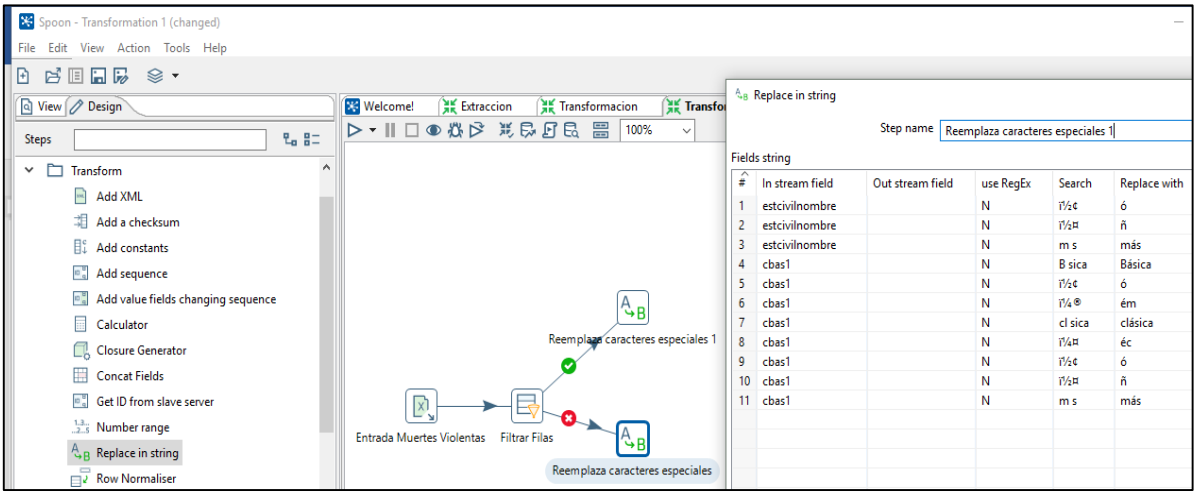


Figura 36: Componente para reemplazar caracteres especiales con PDI
Fuente: Elaboración Propia

El componente "**Reemplaza caracteres especiales 1**", trata los registros filtrados de las filas que no correspondían a los campos, y el componente "**Reemplaza caracteres especiales**", trata los demás registros. Después de haber reemplazado los caracteres especiales, se concatena los campos para dejar los registros en sus campos correspondientes, se debe ir al contenedor "**Transform**", y se arrastra el componente "**Concat Fields**", que se renombra como "**Concatenar campos**", como se ve en la figura 37.

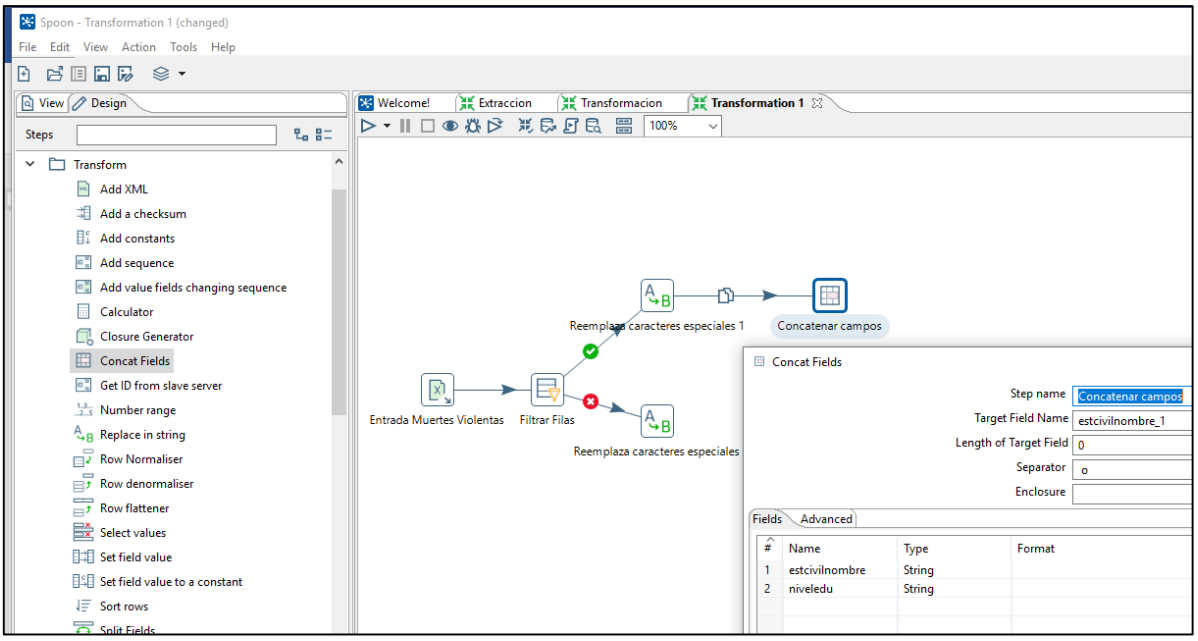


Figura 37: Componente para concatenar los campos con PDI
Fuente: Elaboración Propia

Luego de haber concatenado los campos, se prosigue a renombrar campos, se va al contenedor "**Transform**", y se arrastra el componente "**Set field value**", que se renombra como "**Cambio de valor de campos**", como se puede observar en la figura 38.

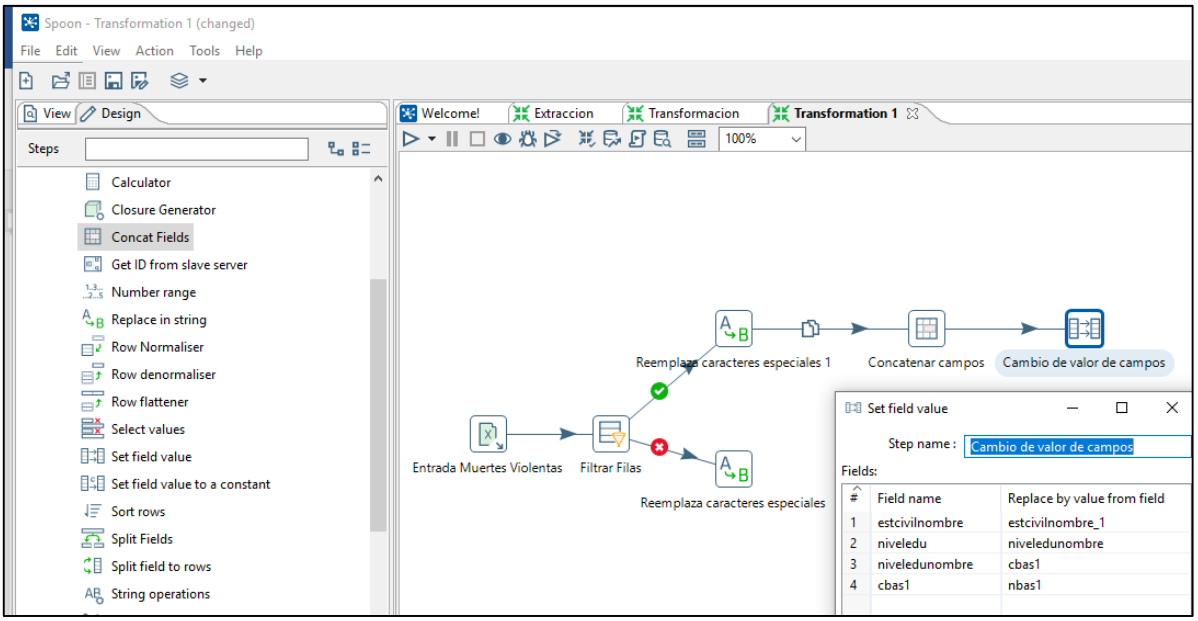
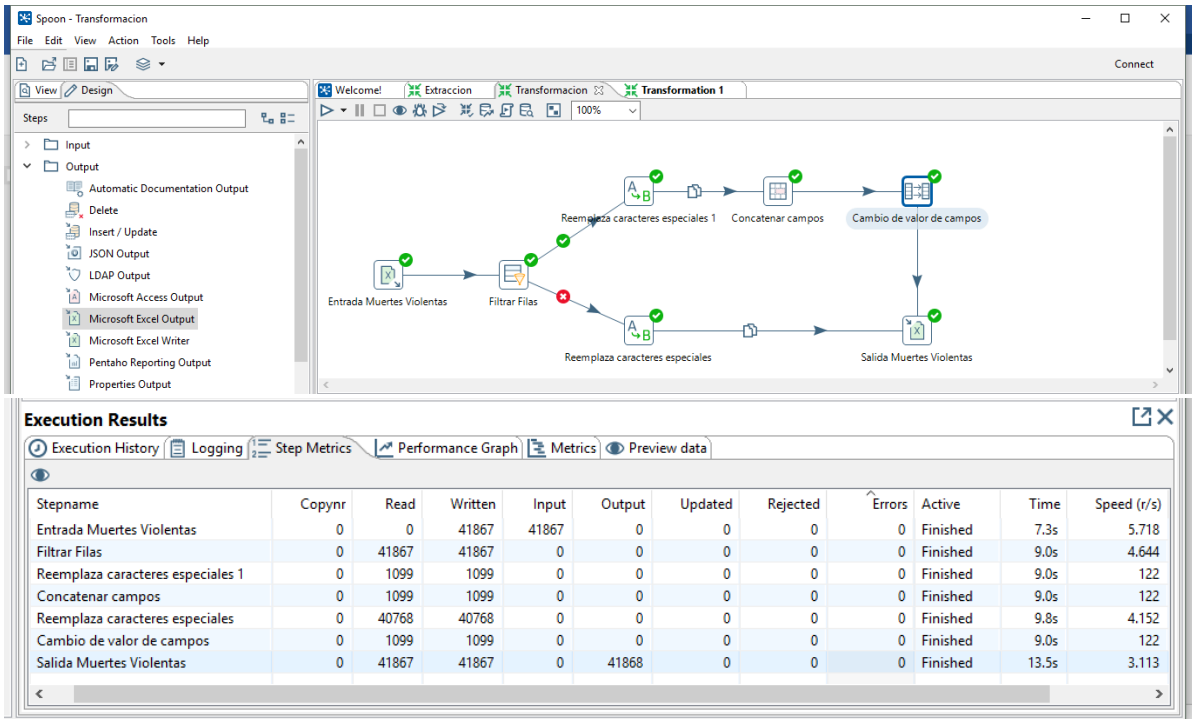


Figura 38: Componente de cambio de valor con PDI
Fuente: Elaboración Propia

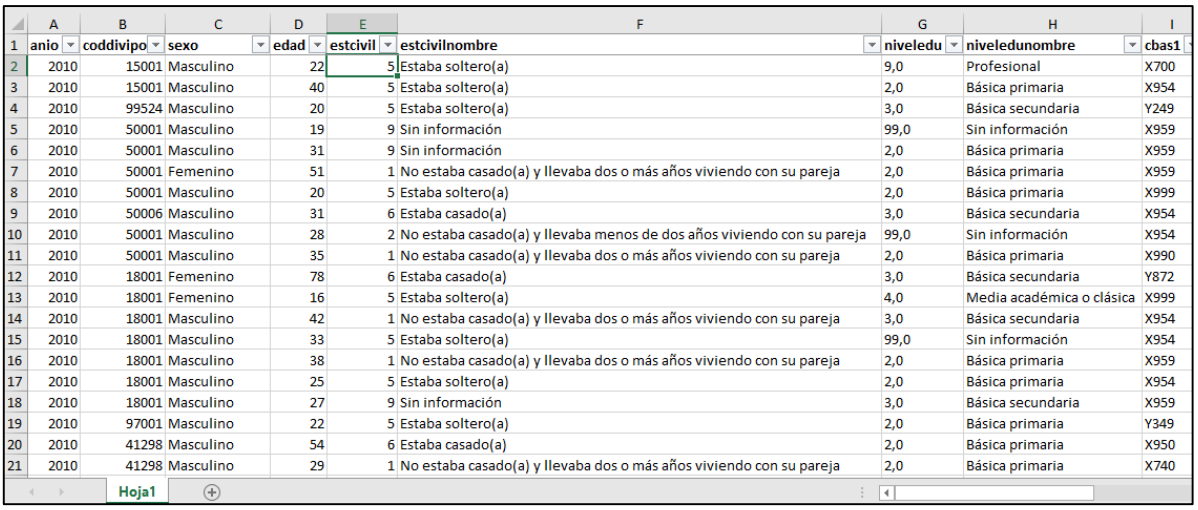
Por último, se carga los datos en un archivo de Excel, se debe ir al contenedor "**Output**", y arrastrar el componente "**Microsoft Excel Output**", que se renombra como "**Salida Muertes Violentas**", que va a contener los datos después de haber realizado el proceso de transformación y de ejecutar el PDI, como se ilustra en la figura 39.



Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)
Entrada Muertes Violentas	0	0	41867	41867	0	0	0	0	Finished	7.3s	5.718
Filtrar Filas	0	41867	41867	0	0	0	0	0	Finished	9.0s	4.644
Reemplaza caracteres especiales 1	0	1099	1099	0	0	0	0	0	Finished	9.0s	122
Concatenar campos	0	1099	1099	0	0	0	0	0	Finished	9.0s	122
Reemplaza caracteres especiales	0	40768	40768	0	0	0	0	0	Finished	9.8s	4.152
Cambio de valor de campos	0	1099	1099	0	0	0	0	0	Finished	9.0s	122
Salida Muertes Violentas	0	41867	41867	0	41868	0	0	0	Finished	13.5s	3.113

Figura 39: Proceso de transformación de datos de muertes violentas en PDI
Fuente: Elaboración Propia

Como resultado de haber ejecutado el programa, arrojó que todo el proceso de transformación duró menos de 13.5 segundos, de los cuales 7.5 segundos se tardó en extraer los datos del archivo de excel a una velocidad de 5.718 registros por segundo, continuando con el filtro de filas que tardó 9.0 segundos a una velocidad de 4.644 registros por segundo, con el reemplazo de caracteres especiales1 que tardó 9.0 segundos a una velocidad de 122 registros por segundo, con el proceso de concatenar campos que tardó 9.0 segundos a una velocidad de 122 registros por segundo, continuando con el reemplazo de caracteres especiales que tardó 9.8 segundos a una velocidad de 4.152 registros por segundo, después con el cambio de valor de campos que tardó 9.0 segundos a una velocidad de 122 registros por segundo y finalizando con la salida de los datos de muertes violentas que tardo 13.5 segundos a una velocidad de 3.113 registros por segundo. Por último, en la figura 40, se puede ver el resultado del archivo de Excel transformado, que no posee caracteres especiales, y se puede ver que los registros ya están en sus campos correspondientes, gracias a la transformación hecha por PDI al archivo.



	A	B	C	D	E	F	G	H	I
1	anio	coddivipo	sexo	edad	estcivil	estcivilnombre	niveledu	niveledunombre	cbas1
2	2010	15001	Masculino	22	5	Estaba soltero(a)	9,0	Profesional	X700
3	2010	15001	Masculino	40	5	Estaba soltero(a)	2,0	Básica primaria	X954
4	2010	99524	Masculino	20	5	Estaba soltero(a)	3,0	Básica secundaria	Y249
5	2010	50001	Masculino	19	9	Sin información	99,0	Sin información	X959
6	2010	50001	Masculino	31	9	Sin información	2,0	Básica primaria	X959
7	2010	50001	Femenino	51	1	No estaba casado(a) y llevaba dos o más años viviendo con su pareja	2,0	Básica primaria	X959
8	2010	50001	Masculino	20	5	Estaba soltero(a)	2,0	Básica primaria	X999
9	2010	50006	Masculino	31	6	Estaba casado(a)	3,0	Básica secundaria	X954
10	2010	50001	Masculino	28	2	No estaba casado(a) y llevaba menos de dos años viviendo con su pareja	99,0	Sin información	X954
11	2010	50001	Masculino	35	1	No estaba casado(a) y llevaba dos o más años viviendo con su pareja	2,0	Básica primaria	X990
12	2010	18001	Femenino	78	6	Estaba casado(a)	3,0	Básica secundaria	Y872
13	2010	18001	Femenino	16	5	Estaba soltero(a)	4,0	Media académica o clásica	X999
14	2010	18001	Masculino	42	1	No estaba casado(a) y llevaba dos o más años viviendo con su pareja	3,0	Básica secundaria	X954
15	2010	18001	Masculino	33	5	Estaba soltero(a)	99,0	Sin información	X954
16	2010	18001	Masculino	38	1	No estaba casado(a) y llevaba dos o más años viviendo con su pareja	2,0	Básica primaria	X959
17	2010	18001	Masculino	25	5	Estaba soltero(a)	2,0	Básica primaria	X954
18	2010	18001	Masculino	27	9	Sin información	3,0	Básica secundaria	X959
19	2010	97001	Masculino	22	5	Estaba soltero(a)	2,0	Básica primaria	Y349
20	2010	41298	Masculino	54	6	Estaba casado(a)	2,0	Básica primaria	X950
21	2010	41298	Masculino	29	1	No estaba casado(a) y llevaba dos o más años viviendo con su pareja	2,0	Básica primaria	X740

Hoja1

Figura 40: Datos de muertes violentas transformados con PDI
Fuente: Elaboración Propia

En conclusión, se debe tener en cuenta que PDI posee muchos tipos de transformaciones, de los cuales en este ejemplo no se utilizaron, que son muy útiles dependiendo del problema que desee solucionar. Se pueden transformar datos no estructurados a estructurados, datos semi-estructurados a estructurados. Luego de la ejecución, en cuanto a la velocidad de lectura y escritura de datos, tiempo que tarda en ejecución del proceso y facilidad de transformación de datos, se puede decir que PDI es una herramienta que posee buena velocidad de lectura y escritura de datos, y que no requiere mucho tiempo para ejecutar las transformaciones, en cuanto a la facilidad, se debe mencionar que en la primera interacción con la herramienta esta podría no ser tan fácil e intuitiva, porque posee un grado de dificultad al tratar de transformar un conjunto de datos, dependiendo de la procedencia de los datos, es decir, si se van a tratar datos estructurados, semi-estructurados o no estructurados.

6.2 Transformación de los datos de muertes violentas con TDI

Para realizar el proceso de transformación en TDI de los datos de muertes violentas, primero se debe ir a el contenedor “**Job Designs**”, y crear un “**Job**”, como se puede ver en la figura 41.

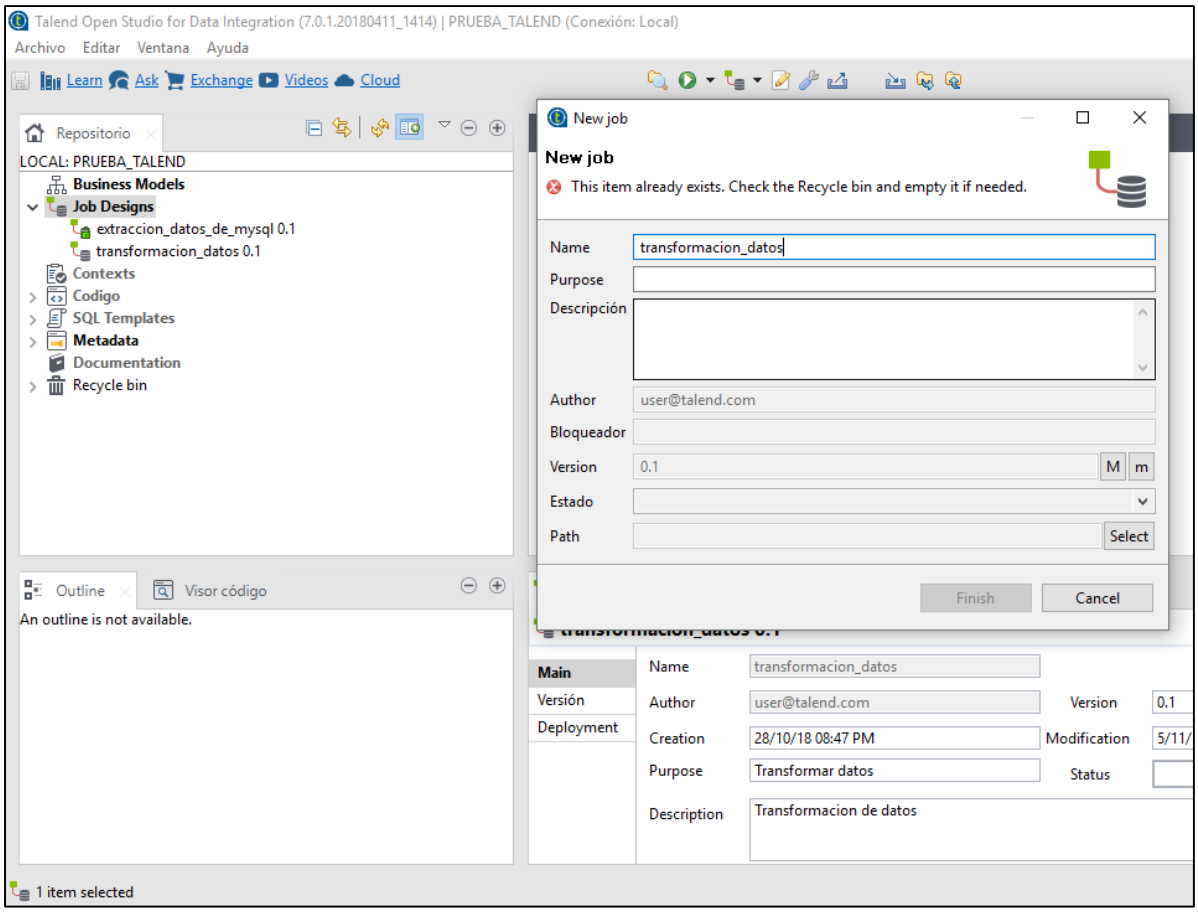


Figura 41: Creación del Job para la transformación de datos con TDI
Fuente: Elaboración Propia

Luego se debe crear el esquema de metadatos, para ello se debe ir al contenedor "**Metadata**", y en el componente "**File Excel**", que sirve para crear una conexión con el archivo de muertes violentas, como se ilustra en la figura 42.

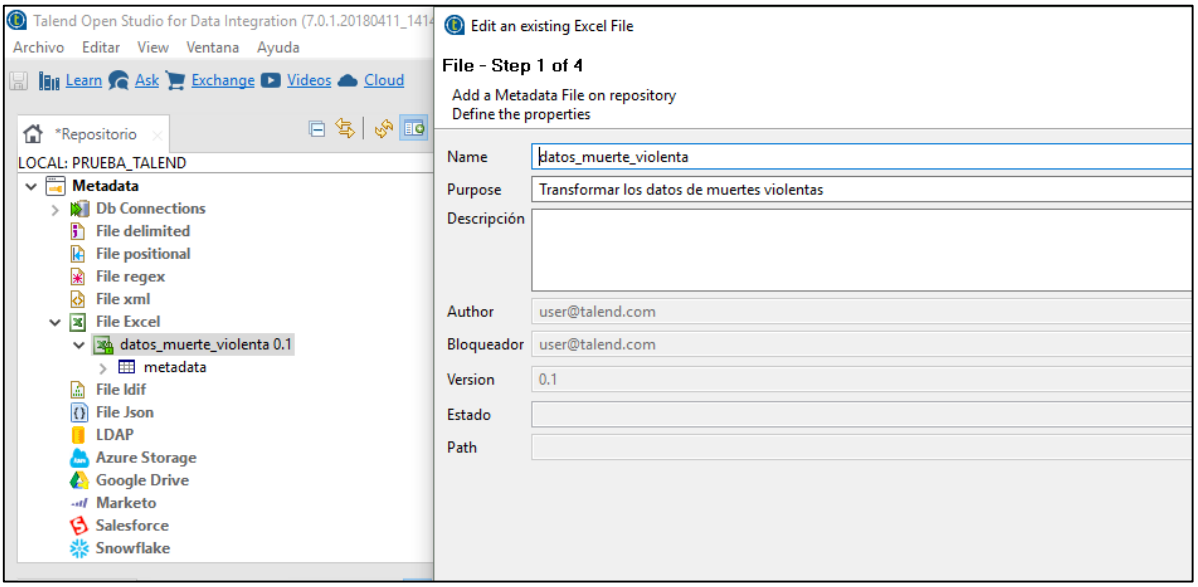


Figura 42: Creación de los metadatos de los datos de muertes violentas con TDI
Fuente: Elaboración Propia

Después de haber creado tanto el Job, como la estructura de metadatos, se prosigue a realizar la transformación de los datos. Primero se deben extraer los datos de muertes violentas del archivo de Excel, y luego se arrastra el componente "**datos_muerte_violenta**", como se ve en la figura 43.

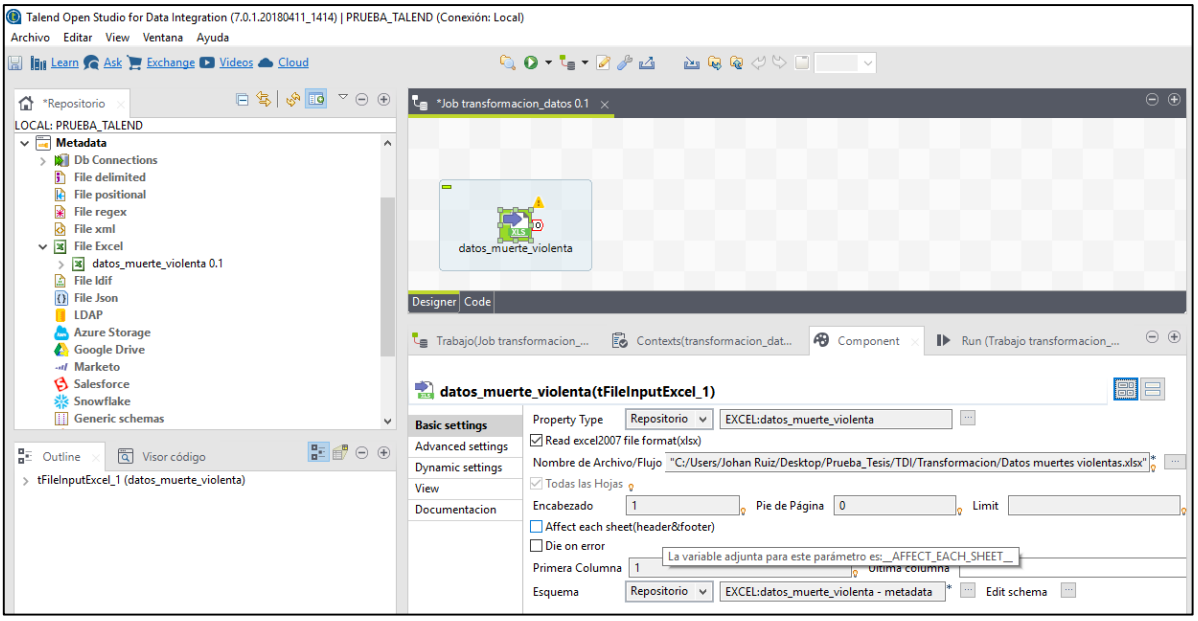


Figura 43: Componente para extraer los datos de muertes violentas con TDI
Fuente: Elaboración Propia

Posteriormente, se prosigue a filtrar las filas que están en campos que no corresponden a su tipo de datos, para eso debe ir al contenedor "**Processing**", y arrastrar el componente "**tFilterRow**", que se renombra como "**Filtrar Filas**", como se puede observar en la figura 44.

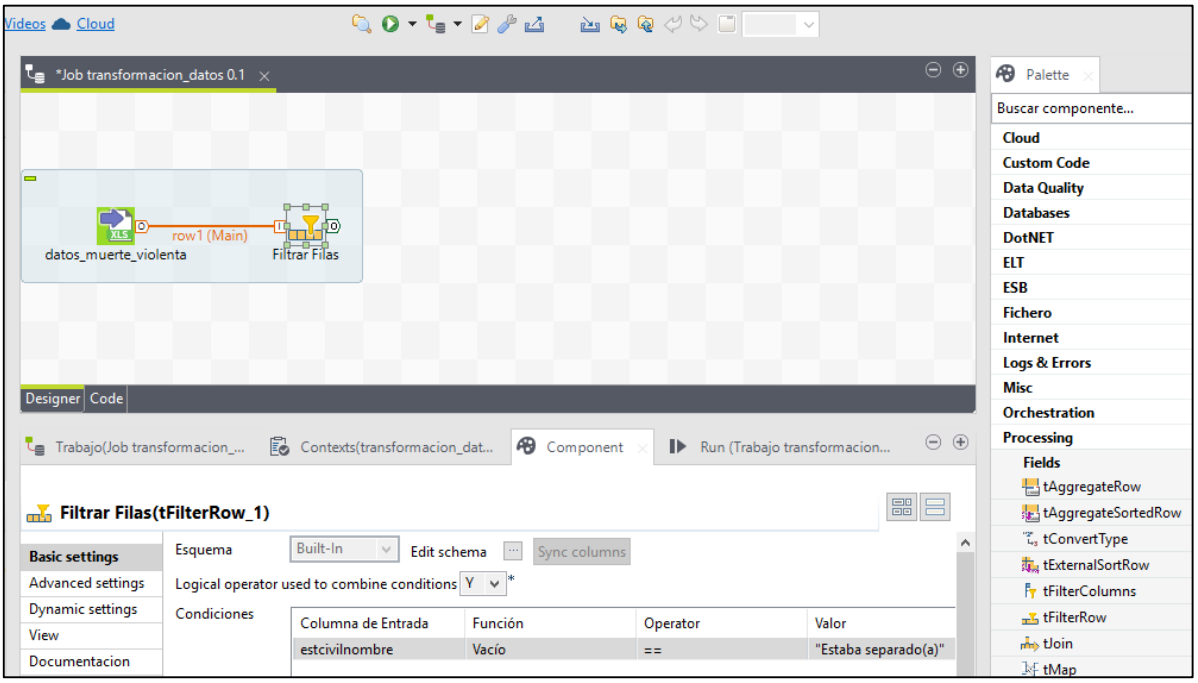


Figura 44: Componente para filtrar filas con TDI
Fuente: Elaboración Propia

A continuación, se reemplaza los caracteres especiales que poseen algunos registros, va al contenedor “**Processing**”, y se arrastra el componente “t**Replace**”, que se renombra como “**Reemplaza caracteres especiales**” y el otro como “**Reemplaza caracteres especiales 1**”, como se ilustra en la figura 45.

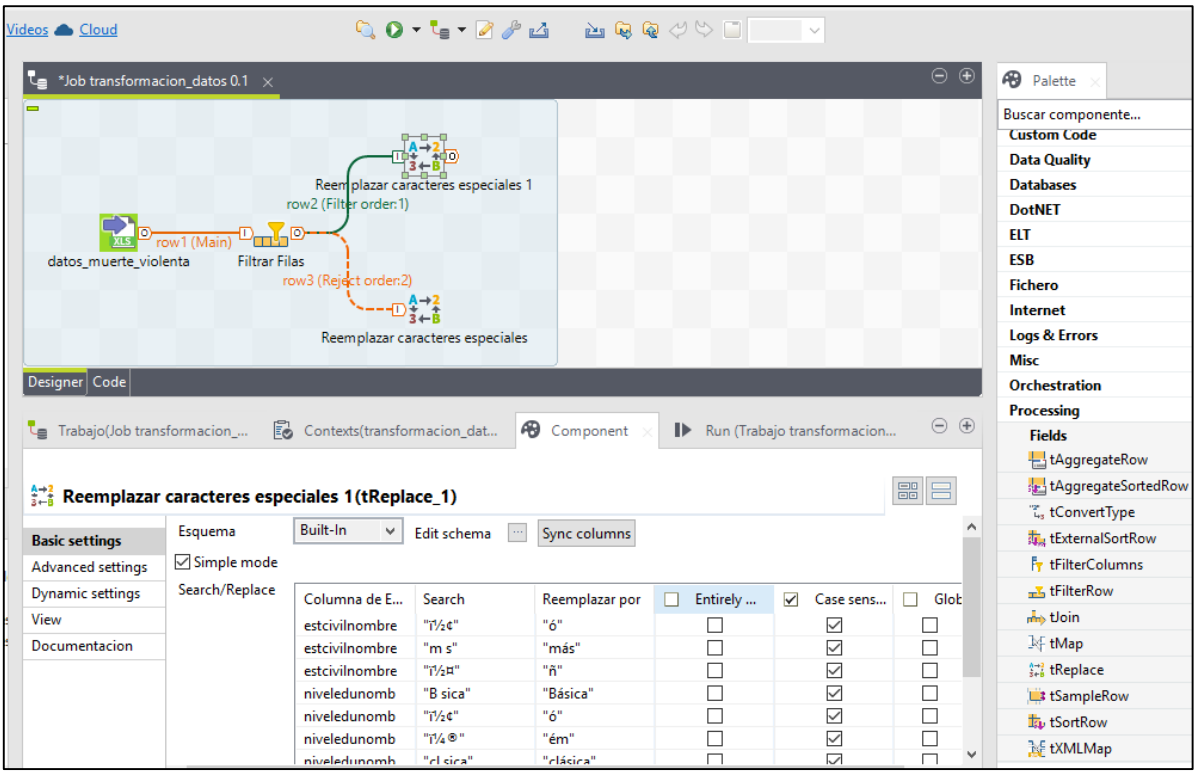


Figura 45: Componente para reemplazar caracteres especiales con TDI
Fuente: Elaboración Propia

Luego de haber reemplazado los caracteres especiales, se concatena los campos para dejar los registros en sus campos correspondientes, debe ir al contenedor “**Processing**”, y se arrastra el componente “**tMap**”, que permite realizar muchos procesos al tiempo como uniones, transformaciones, filtros de columnas, filtros de filas, y múltiples salidas entre otras cosas, que se renombra como “**Union tMap**”, como se ve en la figura 46 y en la figura 47 se observa cómo se concatenan los campos.

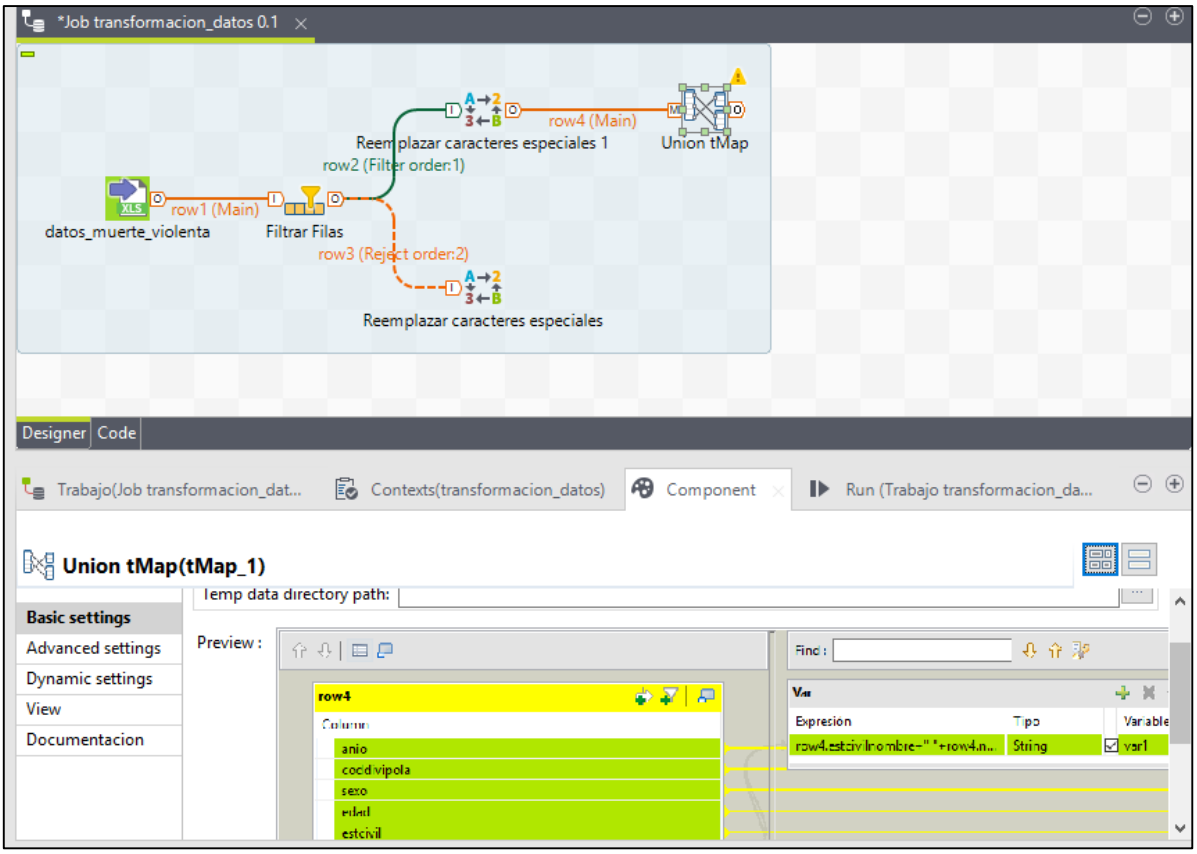


Figura 46: Componente para concatenar los campos con TDI
Fuente: Elaboración Propia

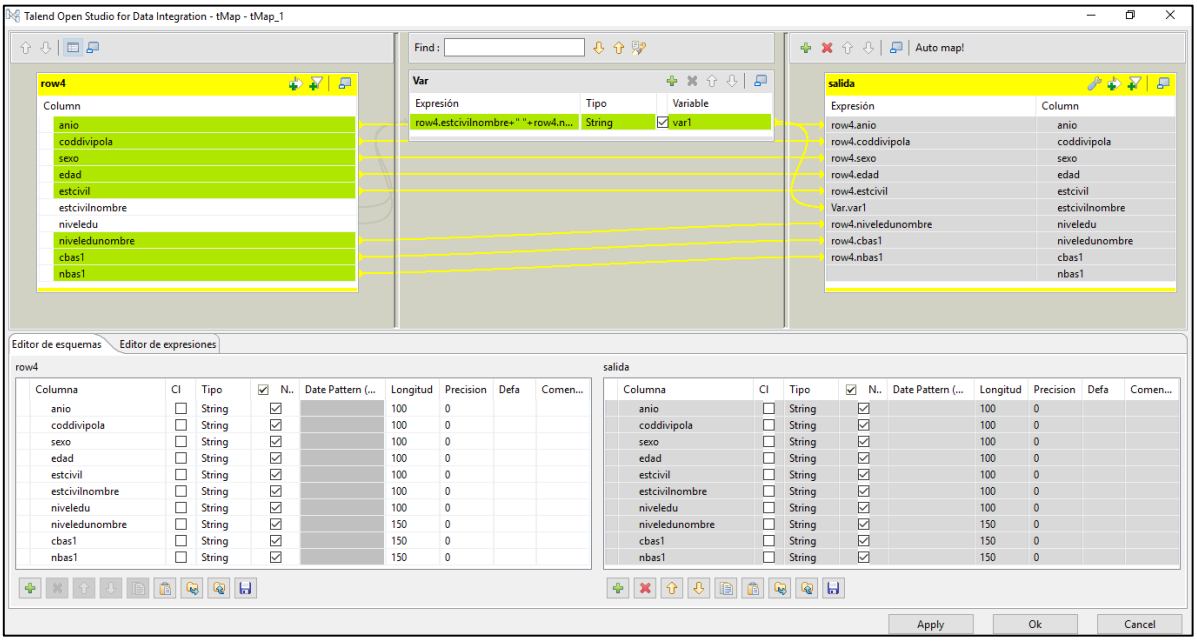


Figura 47: Componente tMap para realizar la concatenación con TDI
Fuente: Elaboración Propia

Por último, se carga los datos en un archivo de Excel, se debe ir al contenedor "**Fichero**", y arrastrar el componente "**tFileOutputExcel**", que se renombra como "**Salida Muertes Violentas**", que va a contener los datos, después de haber realizado el proceso de transformación y de ejecutar el programa, como se puede ver en la figura 48.

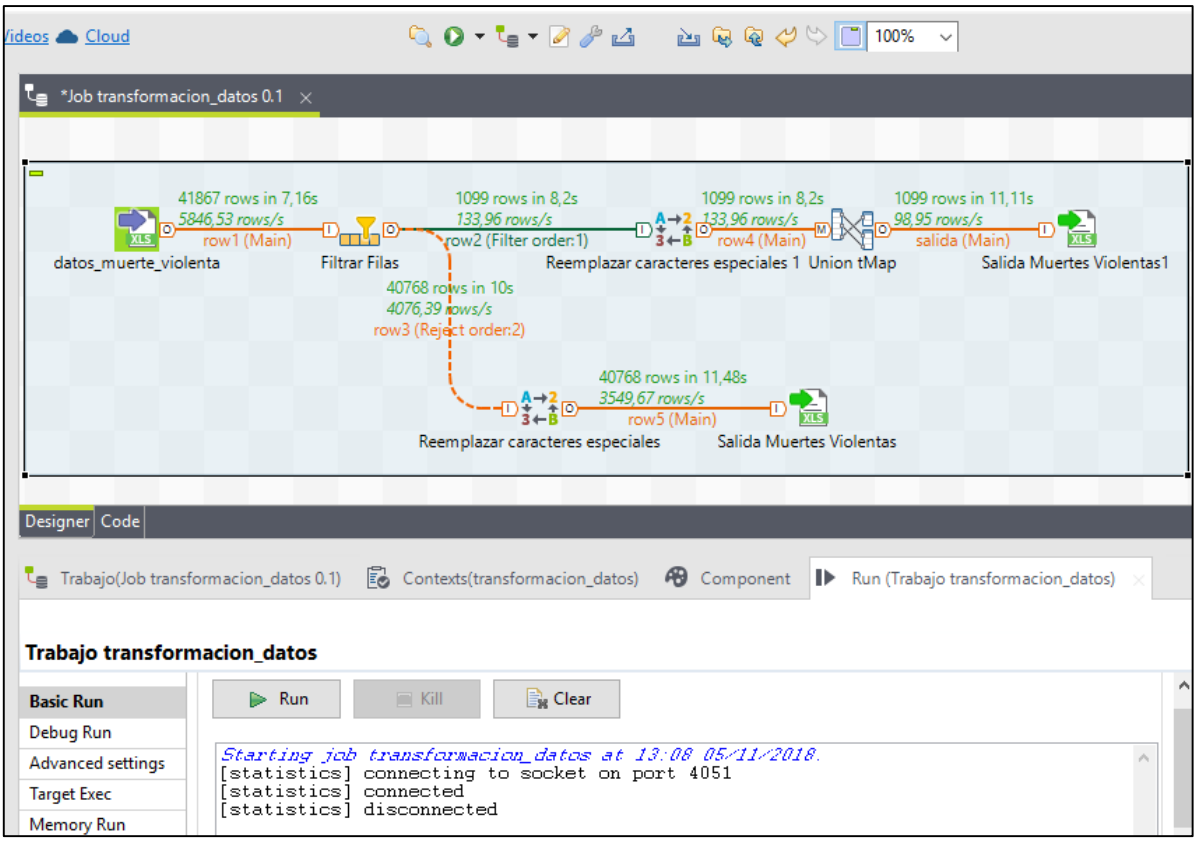
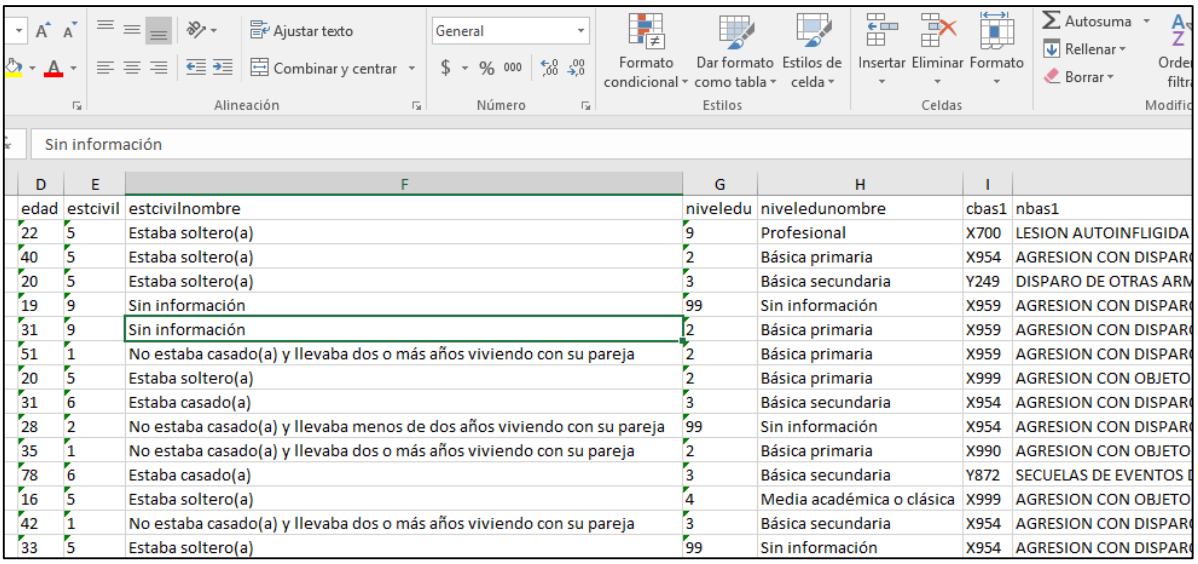


Figura 48: Proceso de transformación de datos de muertes violentas con TDI
Fuente: Elaboración Propia

Como resultado de haber ejecutado el programa, se observa que todo el proceso de transformación duró menos de 11.48 segundos, de los cuales 7.16 segundos se tardó en extraer los datos del archivo de excel a una velocidad de 5.846 registros por segundo, continuando con el filtro de filas que tardó 8.2 segundos a una velocidad de 133 registros por segundo, con el reemplazo de caracteres especiales1 que tardó 8.2 segundos a una velocidad de 133 registros por segundo, con el proceso de concatenar campos que tardó 11.11 segundos a una velocidad de 98 registros por segundo, continuando con el reemplazo de caracteres especiales que tardó 10 segundos a una velocidad de 4.076 registros por segundo, y finalizando con la salida de los datos de muertes violentas que tardo 11.48 segundos a una velocidad 3.549 registros por segundo. Por último, en la figura 49, se puede ver el resultado del archivo de Excel transformado, que no posee caracteres especiales, y se puede ver que los registros ya están en sus campos correspondientes, gracias a la transformación hecha por TDI a los datos.



D	E	F	G	H	I	
edad	estcivil	estcivilnombre	niveledu	niveledunombre	cbas1	nbas1
22	5	Estaba soltero(a)	9	Profesional	X700	LESION AUTOINFLIGIDA
40	5	Estaba soltero(a)	2	Básica primaria	X954	AGRESION CON DISPARO
20	5	Estaba soltero(a)	3	Básica secundaria	Y249	DISPARO DE OTRAS ARM
19	9	Sin información	99	Sin información	X959	AGRESION CON DISPARO
31	9	Sin información	2	Básica primaria	X959	AGRESION CON DISPARO
51	1	No estaba casado(a) y llevaba dos o más años viviendo con su pareja	2	Básica primaria	X959	AGRESION CON DISPARO
20	5	Estaba soltero(a)	2	Básica primaria	X999	AGRESION CON OBJETO
31	6	Estaba casado(a)	3	Básica secundaria	X954	AGRESION CON DISPARO
28	2	No estaba casado(a) y llevaba menos de dos años viviendo con su pareja	99	Sin información	X954	AGRESION CON DISPARO
35	1	No estaba casado(a) y llevaba dos o más años viviendo con su pareja	2	Básica primaria	X990	AGRESION CON OBJETO
78	6	Estaba casado(a)	3	Básica secundaria	Y872	SECUELAS DE EVENTOS
16	5	Estaba soltero(a)	4	Media académica o clásica	X999	AGRESION CON OBJETO
42	1	No estaba casado(a) y llevaba dos o más años viviendo con su pareja	3	Básica secundaria	X954	AGRESION CON DISPARO
33	5	Estaba soltero(a)	99	Sin información	X954	AGRESION CON DISPARO

Figura 49: Datos de muertes violentas transformados con TDI
Fuente: Elaboración Propia

En conclusión, se debe tener en cuenta que TDI posee muchos tipos de transformaciones, de los cuales en este ejemplo no se utilizaron todas, que son muy útiles dependiendo del problema que desee solucionar. Se pueden transformar datos no estructurados a estructurados, datos semi-estructurados a estructurados. Luego de la ejecución, en cuanto a la velocidad de lectura y escritura de datos, tiempo que tarda en ejecución del proceso y facilidad de transformación de datos, se puede decir que TDI es una herramienta que posee buena velocidad de lectura y escritura de datos, y que no requiere mucho tiempo para ejecutar las transformaciones, en cuanto a la facilidad, se debe mencionar que la primera interacción con la herramienta podría no ser tan fácil e intuitiva, porque posee un grado de dificultad al tratar de transformar un conjunto de datos, dependiendo de la procedencia de los datos es decir, si se van a tratar datos estructurados, semi-estructurados o no estructurados, también se debe tener en cuenta que el proceso de entrada y salida de datos en TDI, se maneja por estructuras de metadatos.

6.3 Transformación de los datos de muertes violentas con OR

Para realizar el proceso de transformación en OR de los datos de muertes violentas, primero se debe ir al componente “**Este equipo**” y dar clic en el botón “**Elegir archivos**” y seleccionar la ubicación donde está el archivo de muertes violentas como se ve en la figura 50.

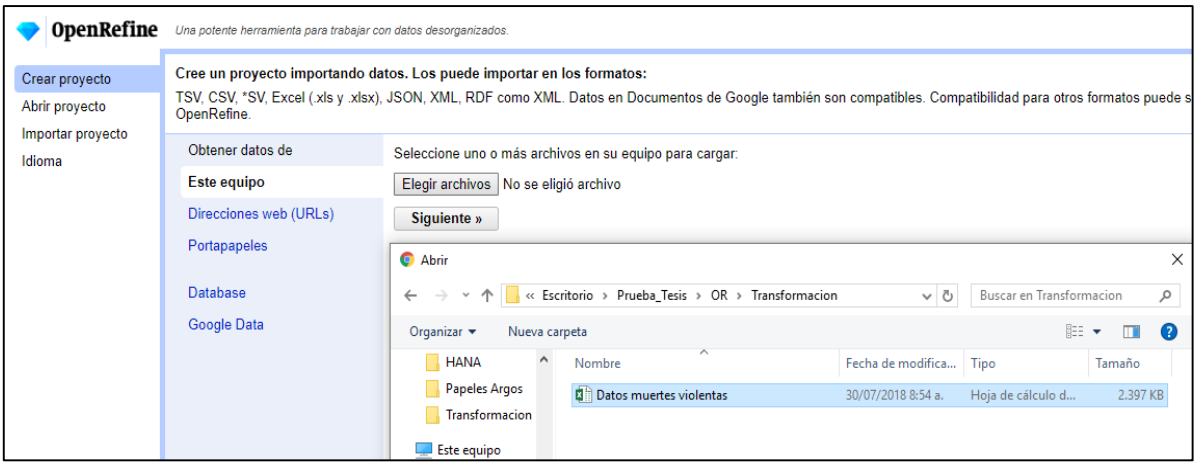


Figura 50: Componente de extracción de datos de muertes violentas con OR
Fuente: Elaboración Propia

Después se prosigue a crear un proyecto con el nombre de "**Transformación**", como se puede ver en la figura 51. Que permite realizar la transformación de los datos.

Figura 51: Creación del proyecto transformación con OR
Fuente: Elaboración Propia

Luego de extraer los datos y de crear el proyecto, se inicia el proceso de transformación. Primero se concatena las filas, es decir en este caso se quiere concatenar "**Estaba separado (a)**" y "**divorciado (a)**" en una sola columna, porque están en columnas separadas y hacen que los registros estén movidos hacía la derecha en otros campos, para ello se va al contendor "**Facetas**" y da clic a "**Faceta de texto**", que abre una ventana al lado izquierdo, que permite concatenar los registros de los campos, como se ilustra en la figura 52.

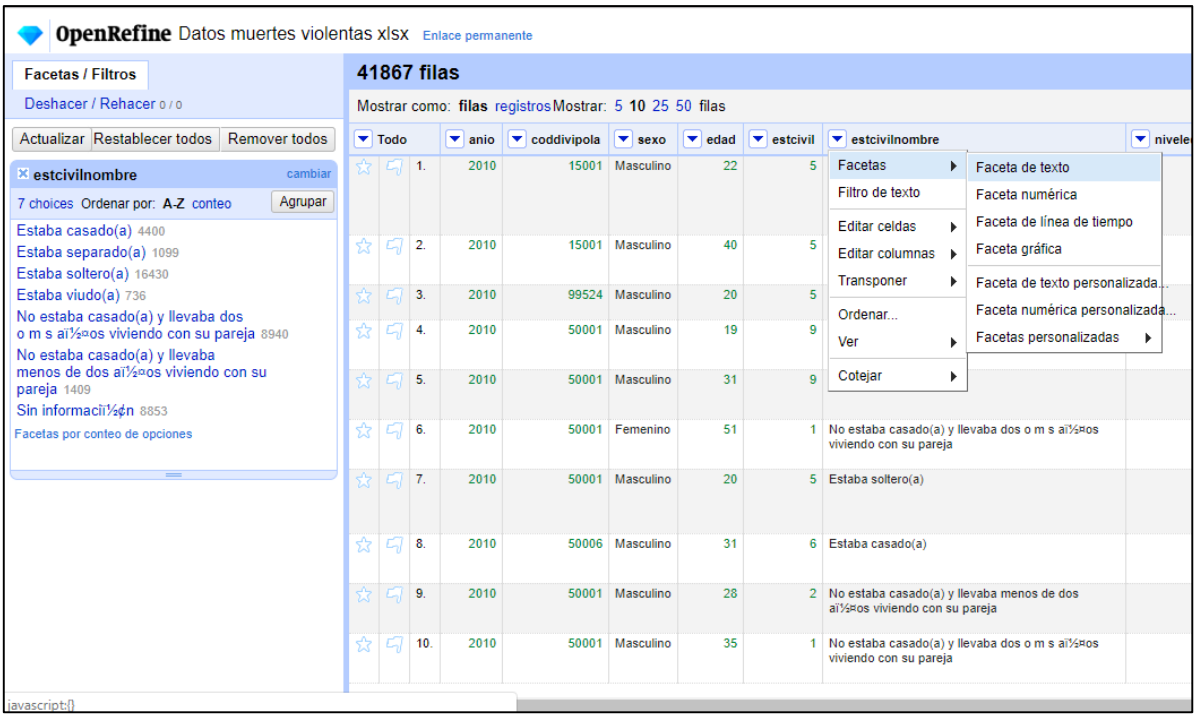


Figura 52: Componente para concatenar los registro con OR
Fuente: Elaboración Propia

Luego de abrir el panel, la figura 53, muestra cómo se concatena los valores correspondientes a "**Estaba separado**" y "**divorciado (a)**".

Figura 53: Proceso para concatenar los registros con OR
Fuente: Elaboración Propia

Después de concatenar los registros de los campos, se prosigue a reemplazar los caracteres especiales, para ello damos clic en cada texto en la ventana izquierda y reemplazamos los caracteres especiales de las columnas "**estacivilnomrbe**", "**niveledunombre**" y "**cbas1**", como se puede ver en las figuras 54, 55 y 56.

Figura 54: Proceso de reemplazar caracteres especiales del campo estcivilnombre con OR
Fuente: Elaboración Propia

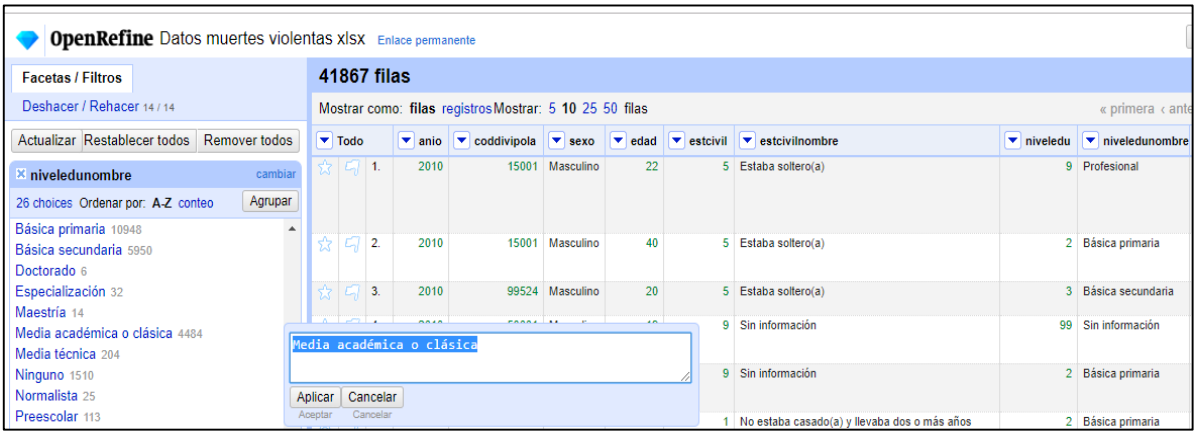


Figura 55: Proceso de reemplazar caracteres especiales del campo niveledunombre con OR
Fuente: Elaboración Propia

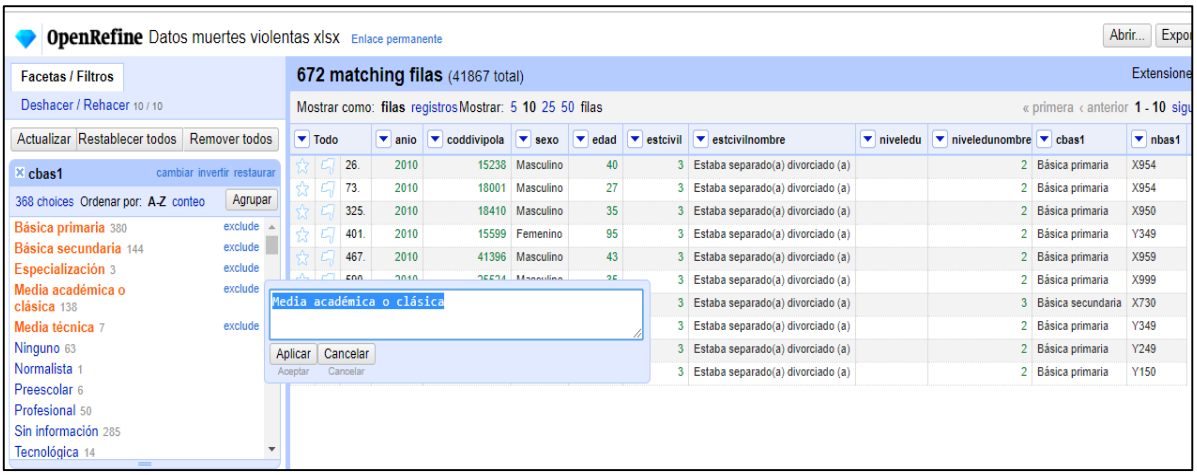


Figura 56: Proceso de reemplazar caracteres especiales del campo cbas1 con OR
Fuente: Elaboración Propia

Luego de reemplazar los caracteres especiales, solo se deben correr los registros que están en otros campos que no corresponden como se puede notar en la figura 56, que requiere mover los registros del campo "**niveledunombre** " a "**niveledu**", los del campos "**cbas1**" a "**niveledunombre**" y los del campo "**nbas1**" a "**cbas1**", para sus campos correspondientes, para eso se ubica en cada campo, en este caso "**niveledu**", y después va al componente

"**Editar celdas**", y da clic a "**Transformar…**", donde se debe agregar el código "**cell.niveledunombre**", que permite copiar los registros del campo "**niveledunombre**", al campo "**niveledu**", como se ve en la figura 57.

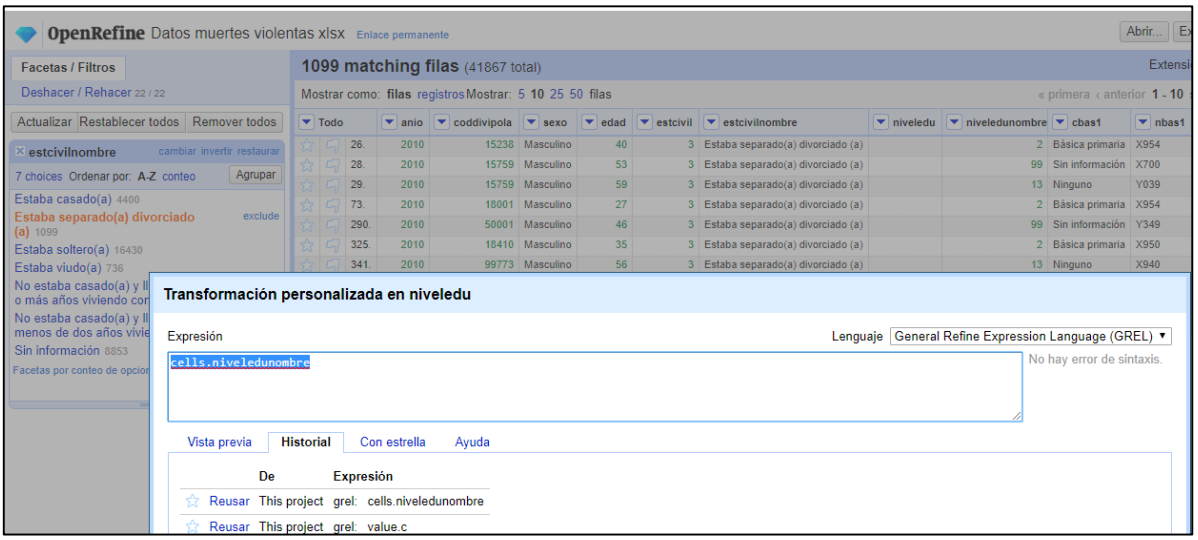


Figura 57: Proceso de mover los registros del campos niveledunombre a niveledu con OR
Fuente: Elaboración Propia

Y prosiguiendo con los demás campos, y al final de mover los registros podemos observar que los registros quedan ubicados en sus campos correspondientes, como se ilustra en la figura 58.

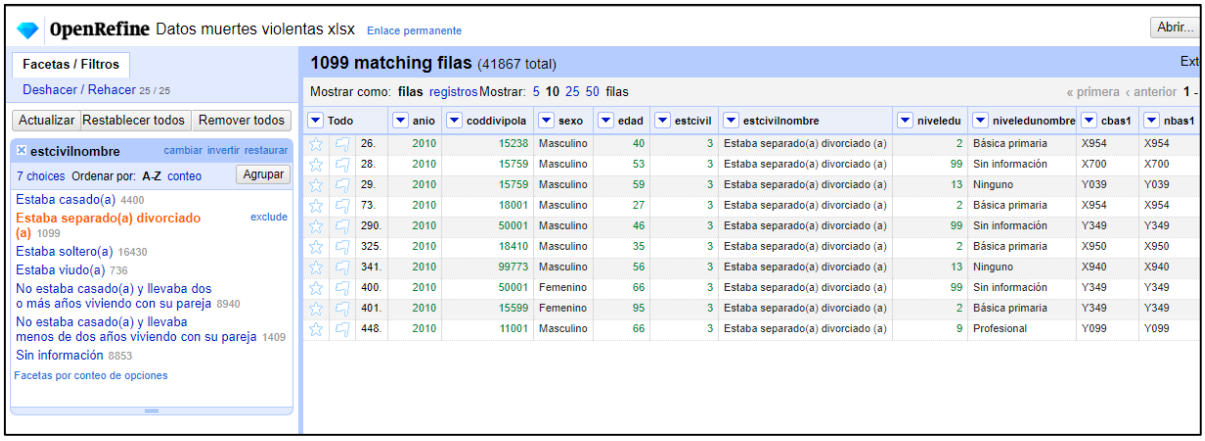


Figura 58: Proceso final de mover los registros a los campos correspondientes con OR
Fuente: Elaboración Propia

Por último, en la figura 59 se observa todo el proceso de transformación, que consta de 25 pasos, y cuyo resultado se ve en la figura 60, donde se puede notar que los datos transformados no tienen problemas de caracteres especiales y los registros quedaron bien ubicados en sus campos correspondientes.

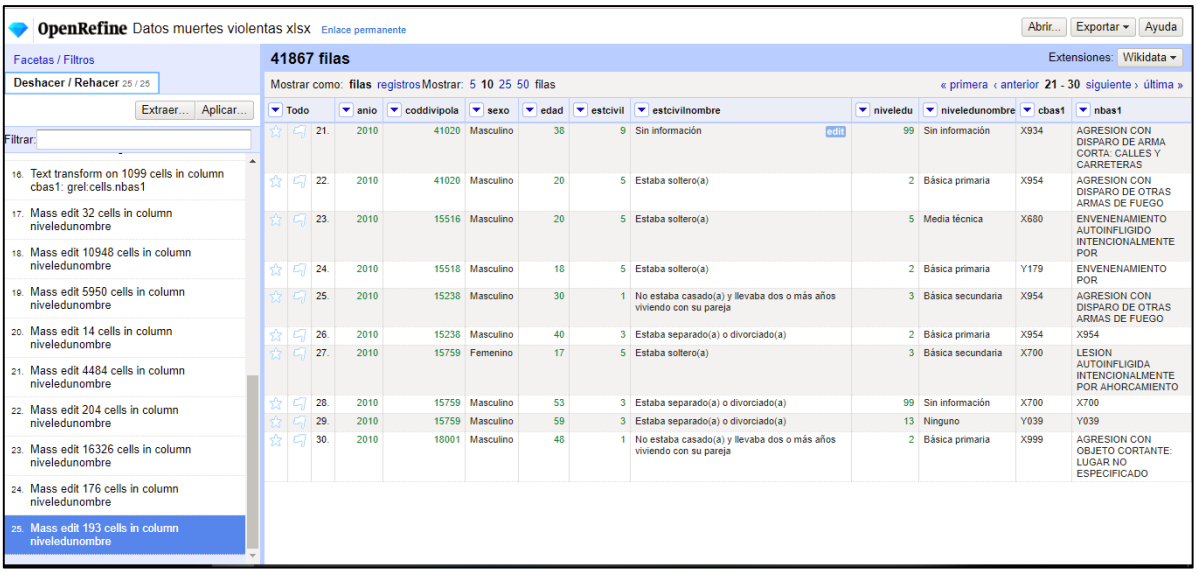
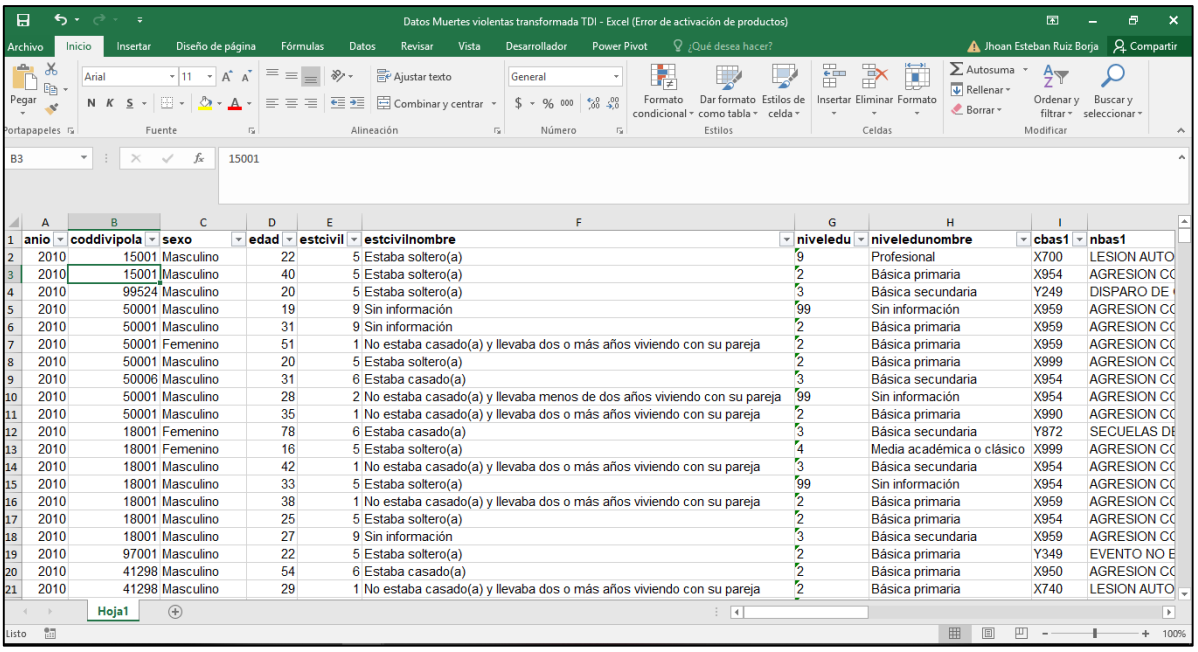


Figura 59: Proceso de transformación de datos de OR
Fuente: Elaboración Propia

En cuanto a la velocidad de lectura y escritura de datos, tiempo que tarda en ejecución del proceso y facilidad de la transformación de datos en OR, en comparación con TDI y PDI, fue rápida, permitiendo transformar datos en el momento que se requiera, teniendo en cuenta que posee la desventaja de que no se puede automatizar procesos, ni tampoco llevar los registros históricos de lectura y escritura de datos. Tenga en cuenta que OR tiene una desventaja de capacidad de procesamiento, ya que requiere una buena máquina para trabajar con gran cantidad de datos. Otra desventaja es que se puede procesar un archivo por proyecto, es decir si son varios archivos se deben abrir varios proyectos, pero se pueden unir si tiene un atributo en común.

Se debe tener en cuenta que OR posee muchos tipos de transformaciones, los cuales en este ejemplo no se utilizaron todas, que son muy útiles dependiendo del problema que desee solucionar, OR permite transformar datos no estructurados en estructurados, también posee unos algoritmos de búsqueda, entre otras transformaciones útiles.



	A	B	C	D	E	F	G	H	I	
1	anio	coddivipola	sexo	edad	estcivil	estcivilnombre	niveledu	niveledunombre	cbas1	nbas1
2	2010	15001	Masculino	22	5	Estaba soltero(a)	9	Profesional	X700	LESION AUTO
3	2010	15001	Masculino	40	5	Estaba soltero(a)	2	Básica primaria	X954	AGRESION CO
4	2010	99524	Masculino	20	5	Estaba soltero(a)	3	Básica secundaria	Y249	DISPARO DE
5	2010	50001	Masculino	19	9	Sin información	99	Sin información	X959	AGRESION CO
6	2010	50001	Masculino	31	9	Sin información	2	Básica primaria	X959	AGRESION CO
7	2010	50001	Femenino	51	1	No estaba casado(a) y llevaba dos o más años viviendo con su pareja	2	Básica primaria	X959	AGRESION CO
8	2010	50001	Masculino	20	5	Estaba soltero(a)	2	Básica primaria	X999	AGRESION CO
9	2010	50006	Masculino	31	6	Estaba casado(a)	3	Básica secundaria	X954	AGRESION CO
10	2010	50001	Masculino	28	2	No estaba casado(a) y llevaba menos de dos años viviendo con su pareja	99	Sin información	X954	AGRESION CO
11	2010	50001	Masculino	35	1	No estaba casado(a) y llevaba dos o más años viviendo con su pareja	2	Básica primaria	X990	AGRESION CO
12	2010	18001	Femenino	78	6	Estaba casado(a)	3	Básica secundaria	Y872	SECUELAS DE
13	2010	18001	Femenino	16	5	Estaba soltero(a)	4	Media académica o clásico	X999	AGRESION CO
14	2010	18001	Masculino	42	1	No estaba casado(a) y llevaba dos o más años viviendo con su pareja	3	Básica secundaria	X954	AGRESION CO
15	2010	18001	Masculino	33	5	Estaba soltero(a)	99	Sin información	X954	AGRESION CO
16	2010	18001	Masculino	38	1	No estaba casado(a) y llevaba dos o más años viviendo con su pareja	2	Básica primaria	X959	AGRESION CO
17	2010	18001	Masculino	25	5	Estaba soltero(a)	2	Básica primaria	X954	AGRESION CO
18	2010	18001	Masculino	27	9	Sin información	3	Básica secundaria	X959	AGRESION CO
19	2010	97001	Masculino	22	5	Estaba soltero(a)	2	Básica primaria	Y349	EVENTO NO E
20	2010	41298	Masculino	54	6	Estaba casado(a)	2	Básica primaria	X950	AGRESION CO
21	2010	41298	Masculino	29	1	No estaba casado(a) y llevaba dos o más años viviendo con su pareja	2	Básica primaria	X740	LESION AUTO

Figura 60: Datos transformados con TDI
Fuente: Elaboración Propia

En conclusión, OR se destaca por ser una herramienta orientada a la transformación de datos, basado en ocho componentes como los son: Facetas, Filtro de texto, Editar celdas, Editar columnas, Transponer, Ordenar, Ver y Cotejar, donde cada componente de estos tiene subcomponentes que le permiten transformar los datos de una forma adecuada. También posee unos algoritmos, uno de agrupación de celdas similares, que funciona después de dividir las celdas de múltiples valores, donde es posible que se note etiquetas que no siempre tienen la misma ortografía. Por ejemplo, hay Trajes y Traje (diferencias de pluralización). El proceso de encontrar los mismos elementos con una ortografía ligeramente diferente se denomina agrupamiento. Una vez que haya dividido las celdas de valores múltiples, donde luego se puede elegir entre diferentes métodos de agrupación, cada uno de los cuales puede usar varias funciones de similitud. El otro algoritmo es la extracción de entidades nombres, que funciona muy bien para aquellos campos que en su conjunto de datos contienen términos únicos, como nombres de personas, ubicaciones, valores, organizaciones, países u obras de arte. Además de solo extraer los términos, la mayoría de los algoritmos también intentan realizar la desambiguación. Por ejemplo, si el algoritmo encuentra a Washington en un texto, intentará determinar si se menciona la ciudad o la persona. Esto nos evita tener que revisar los términos extraídos.

6.4 Conclusión de transformación de datos con PDI, TDI y OR

Si desea realizar el proceso de transformación de datos de múltiples fuentes, la herramienta ETL, que le puede ahorrar tiempo es TDI, seguido por PDI, aunque OR es adecuado para realizar la transformación, y es más fácil e intuitivo de utilizar que las demás herramientas, caso que puede ser útil para personas que están empezando a limpiar datos, siendo el caso de un solo archivo, ya que si son varios archivos y se desea integrarlos en un archivo final no se recomienda OR, ya que es limitado para este caso. También tener en cuenta que TDI es la herramienta que tiene un poco más de complejidad al interactuar por primera vez con ella, en comparación con las demás. En la tabla 6.1 se compara las herramientas con algunos componentes que los hace diferentes, en cuanto a la transformación de datos.

Proceso / Herramienta		Talend Data Integration (TDI)	Pentaho Data Integration (PDI)	OpenRefine (OR)
Facilidad en manejar transformaciones		Facilidad baja	Facilidad media	Facilidad alta
Transformación y Unión de varios archivos		Si	Si	No
Tiempo de ejecución en la transformación del archivo de Excel (41.867 registros y 10 campos)		Rápido (11.58 Segundos)	Rápido (13.5 Segundos)	No es medible dentro del programa como lo hacen las demás herramientas
Velocidad de transformación del archivo de Excel (41.867 registros y 10 campos)		El más rápido Dependiente de la transformación	EL segundo más rápido Dependiente de la transformación	El menos rápido Dependiente de la transformación
Extraer	**Tiempo que tarda en ejecución**	7.3 segundos	7.16 segundos	No es medible dentro del programa
	Velocidad de extracción	Velocidad de 5.718 registros por segundo	Velocidad de 5.856 registros por segundo	No es medible dentro del programa
Filtrar	**Tiempo que tarda en ejecución**	9.0 segundos	10 segundos	No es medible dentro del programa
	Velocidad de transformación	Velocidad de 4.644 registros por segundo	Velocidad 4.076 registros por segundo	No es medible dentro del programa
Reemplazar caracteres	**Tiempo que tarda en ejecución**	9.8 segundos	11.48 segundos	No es medible dentro del programa
	Velocidad de transformación	Velocidad de 4.152 registros por segundo	Velocidad 3.549 registros por segundo	No es medible dentro del programa
Concatenar	**Tiempo que tarda en ejecución**	9.0 segundos	11.11 segundos	No es medible dentro del programa
	Velocidad de transformación	Velocidad de 122 registros por segundo	Velocidad de 98 registros por segundo	No es medible dentro del programa
Cargar	**Tiempo que tarda en ejecución**	13.5 segundos	11.48 segundos	No es medible dentro del programa
	Velocidad de carga	Velocidad de 3.113 registros por segundo	Velocidad de 3.549 registros por segundo	No es medible dentro del programa
Procesos automatizados		Si	Si	No
Necesita generar un esquema de metadatos para la transformación de datos		Si (Metadatos)	No	No
Múltiples transformaciones en un componente		tMap	No	No
Algoritmos de búsqueda		No	No	Si
Transformación de datos estructurados		Si	Si	Si
Transformación de datos semi-estructurados		Si	Si	Si
Transformación de datos no estructurados		Si	Si	Si

Tabla 6-1: Comparación de las herramientas en la componente de Transformación de datos usando los datos de muertes violentas.

7 Comparación de herramientas en el proceso de carga de datos (L)

Como TDI, PDI y OR tienen muchas fuentes donde pueden cargar los datos como por ejemplo (Bases de datos, archivos planos, archivos de Excel, JSON, XML, entre otros), el ejemplo práctico se realizará extrayendo datos del archivo plano (txt) utilizado en el proceso de extracción de datos, con 50.000 registros y 61 campos, y cargarlos en una tabla de la base de datos MySQL, y en un archivo de Excel donde se busca comparar la velocidad de lectura y escritura de datos, tiempo que tarda en ejecución del proceso y facilidad de carga de datos.

7.1 Carga de datos a una tabla de MySQL con PDI

Para realizar en PDI la cargar de datos a una tabla de MySQL, primero se debe crear una transformación, se debe ir a "**File**", luego a "**New**", y dar clic en la opción "**Transformation**", como se puede ver en la figura 61.

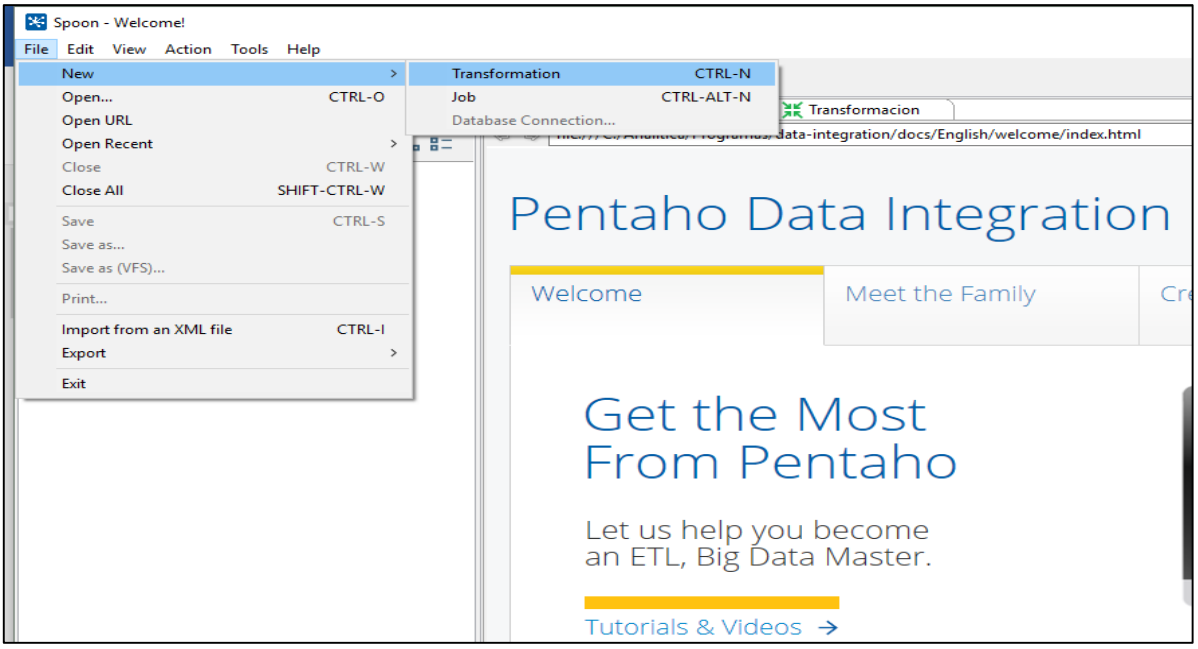


Figura 61: Creación de la transformación para la carga de datos en PDI
Fuente: Elaboración Propia

Después de crear la transformación, se va al contenedor "**Input**", y se arrastra el componente "**Text file input**", que se renombra como "**Entrada de txt**", que sirve para extraer los datos de un archivo plano, como se observa en la figura 62.

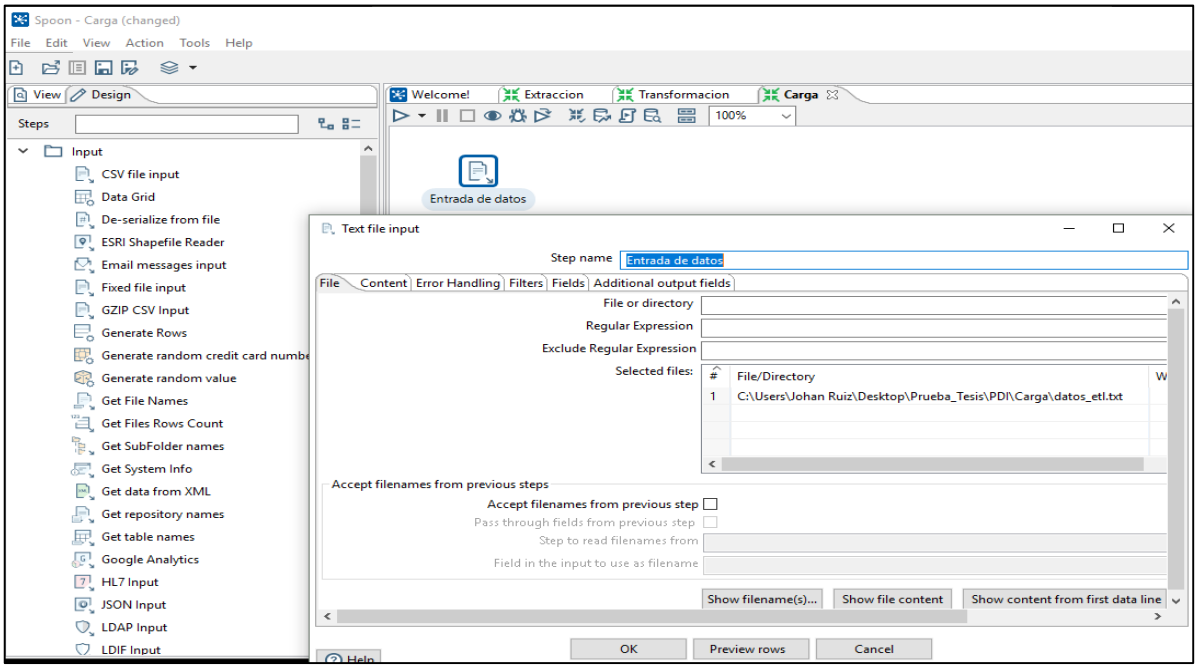


Figura 62: Componente de extracción de datos de un archivo plano con PDI
Fuente: Elaboración Propia

Luego de extraer los datos del archivo plano, se deben cargar en una tabla de MySQL, para ello debe ir al contenedor "**Output**", y se arrastra el componente "**Table output**", que se renombra como "**Salida a tabla de MySQL**", que sirve para poder crear la conexión con la base de datos MySQL y poder cargar los datos, como se ilustra en la figura 63.

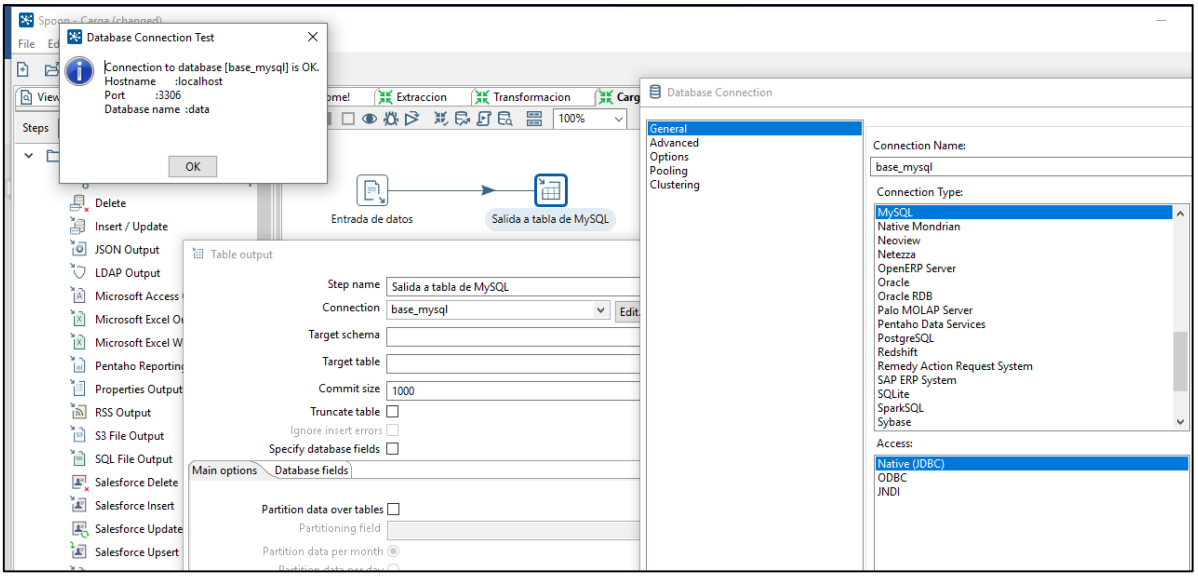


Figura 63: Conexión con la base de datos MySQL con PDI
Fuente: Elaboración Propia

Por último, después de tener el proceso de carga de datos, se prosigue a ejecutar PDI, cuyo resultado se puede ver en la figura 64.

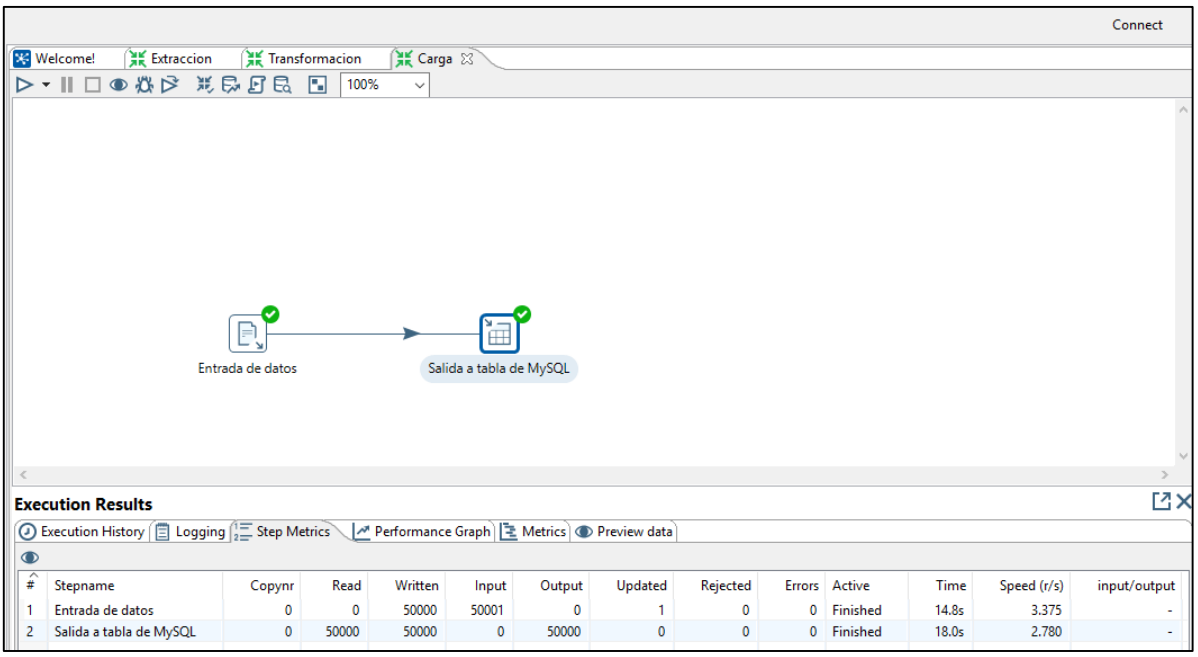
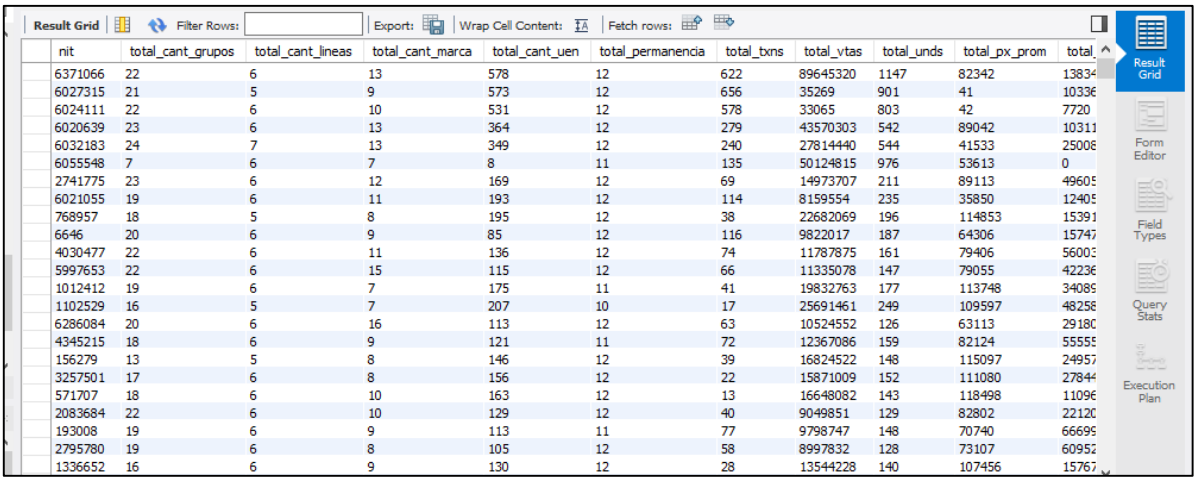


Figura 64: Carga de datos a una tabla de MySQL con PDI
Fuente: Elaboración Propia

En cuanto a velocidad de lectura y escritura de datos, tiempo que tarda en ejecución del proceso y facilidad de carga de datos, el programa arroja como resultado, la extracción de datos de un archivo plano en un tiempo de 14.8 segundos a una velocidad 3.375 registros por

segundo, que se cargan en una tabla de MySQL en un tiempo de 18.0 segundos a una velocidad de 2.780 registros por segundo. Los datos se pueden ver en la figura 65. En cuanto a la facilidad de cargar los datos, se puede decir que no es fácil en la primera interacción con la herramienta, es decir tiene algo de complejidad, tanto en la conexión como creación de la estructura en la que se van a cargar los datos, que deben de ser del mismo tipo, para ello PDI permite crear la estructura directamente por medio de SQL.



nit	total_cant_grupos	total_cant_lineas	total_cant_marca	total_cant_uen	total_permanencia	total_txns	total_vtas	total_unds	total_px_prom	total
6371066	22	6	13	578	12	622	89645320	1147	82342	13834
6027315	21	5	9	573	12	656	35269	901	41	10336
6024111	22	6	10	531	12	578	33065	803	42	7720
6020639	23	6	13	364	12	279	43570303	542	89042	10311
6032183	24	7	13	349	12	240	27814440	544	41533	25008
6055548	7	6	7	8	11	135	50124815	976	53613	0
2741775	23	6	12	169	12	69	14973707	211	89113	49605
6021055	19	6	11	193	12	114	8159554	235	35850	12405
768957	18	5	8	195	12	38	22682069	196	114853	15391
6646	20	6	9	85	12	116	9822017	187	64306	15747
4030477	22	6	11	136	12	74	11787875	161	79406	56003
5997653	22	6	15	115	12	66	11335078	147	79055	42236
1012412	19	6	7	175	11	41	19832763	177	113748	34089
1102529	16	5	7	207	10	17	25691461	249	109597	48258
6286084	20	6	16	113	12	63	10524552	126	63113	29180
4345215	18	6	9	121	11	72	12367086	159	82124	55555
156279	13	5	8	146	12	39	16824522	148	115097	24957
3257501	17	6	8	156	12	22	15871009	152	111080	27844
571707	18	6	10	163	12	13	16648082	143	118498	11096
2083684	22	6	10	129	12	40	9049851	129	82802	22120
193008	19	6	9	113	11	77	9798747	148	70740	66695
2795780	19	6	8	105	12	58	8997832	128	73107	60952
1336652	16	6	9	130	12	28	13544228	140	107456	15767

Figura 65: Datos cargados a la tabla de MySQL con PDI
Fuente: Elaboración Propia

En conclusión PDI, no solo te permite cargar datos a una tabla de MySQL, sino también a muchos más Sistemas Gestores de Base de Datos (Oracle, DB2, PostgreSQL, Teradata, entre otros), permite cargar datos en archivos de Excel (xlsx, xls), archivos planos (csv, txt), archivos JSON, archivos XML, entre otros. Es un gran punto que se debe tener en cuenta a la hora de poder elegir una herramienta, con el proposito de realizar proceso ETL.

7.2 Carga de datos a una tabla de MySQL con TDI

Para realizar en TDI la cargar de datos a una tabla de MySQL, primero se debe crear un Job, para ello se va al contenedor “**Job Designs**” y se crea un “**Job**” que se va a llamar carga, como se ve en la figura 66.

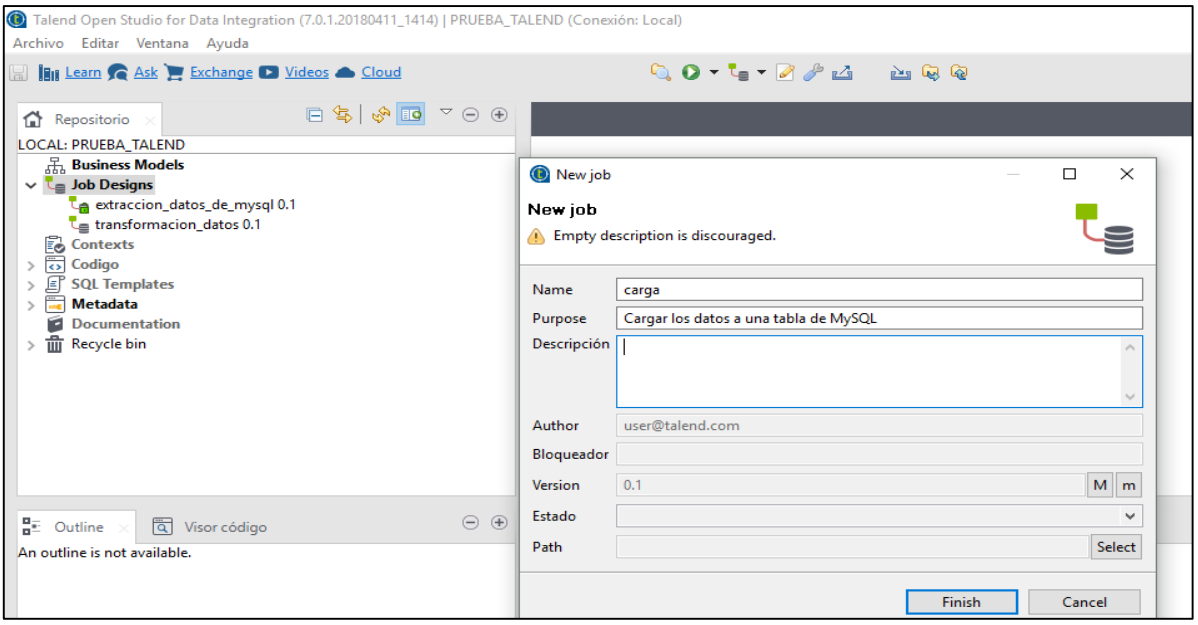


Figura 66: Creación del Job en PDI para la carga de datos
Fuente: Elaboración Propia

Luego de crear el Job, se debe crear los metadatos que establecen la conexión con el archivo plano, para eso se debe ir al contenedor "**Metadata**", y en el componente "**File delimited**", se crea la conexión, como se ilustra en la figura 67.

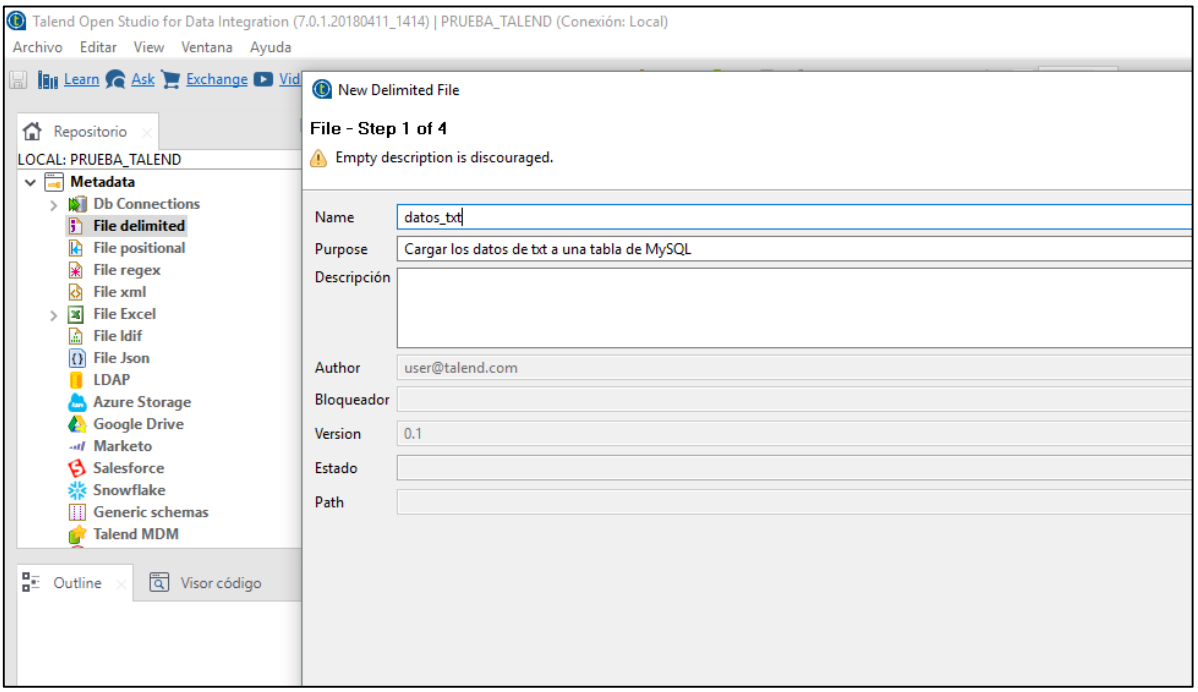


Figura 67: Creación de los metadatos con la conexión al archivo plano con TDI
Fuente: Elaboración Propia

Después de haber creado tanto el Job, como los metadatos, se prosigue a cargar los datos, para eso se debe primero arrastrar el componente "**datos_txt**", que contiene los datos que se van a cargar, como se observa en la figura 68.

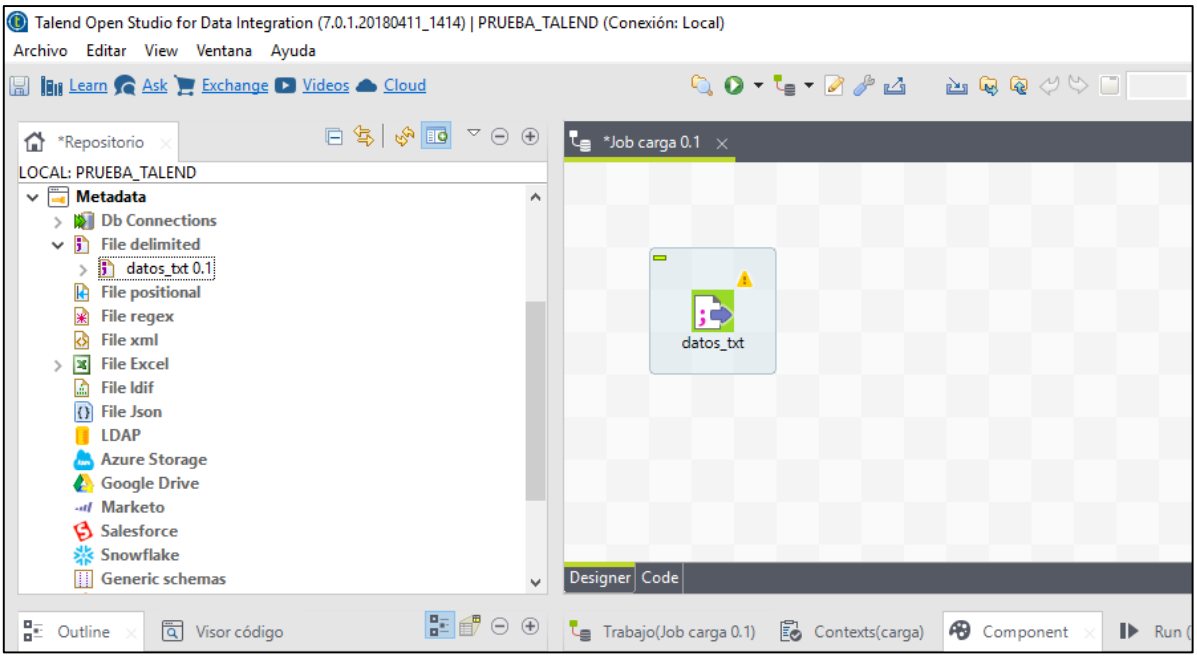


Figura 68: Componente de extracción de datos del archivo plano con TDI
Fuente: Elaboración Propia

Luego de extraer los datos, se debe crear la conexión con la base de datos MySQL, para ello se debe ir al contenedor "**Metadata**", y en el componente "**Db Connections**", se crea la conexión, como se puede ver en la figura 69.

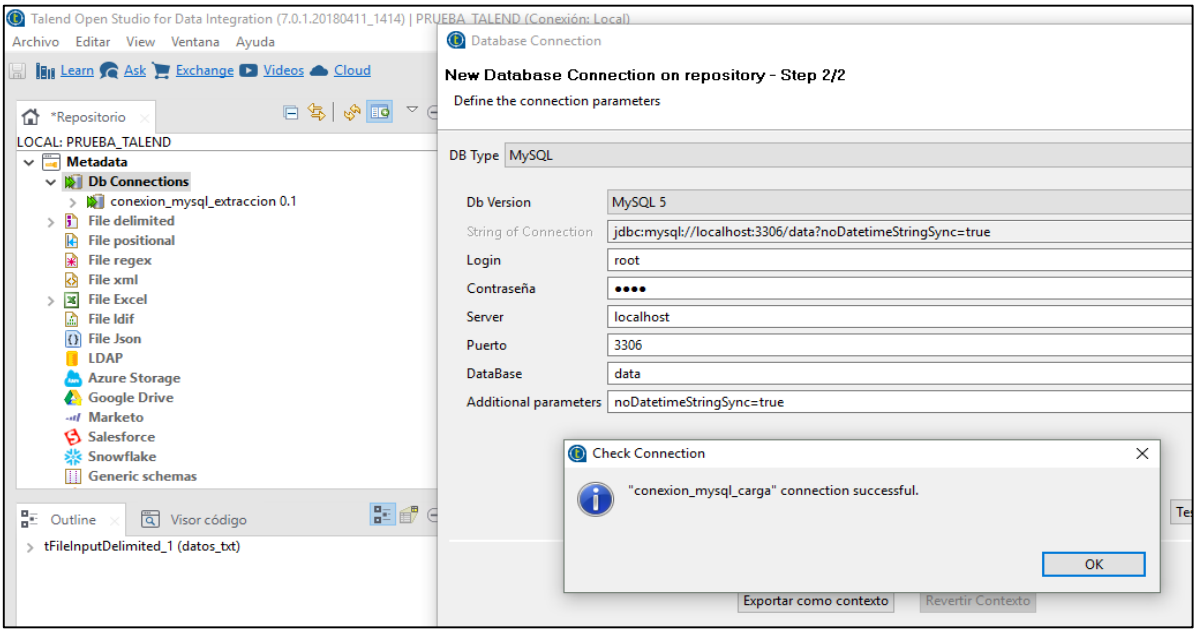


Figura 69: Conexión con la base de datos MySQL para la carga de datos con TDI
Fuente: Elaboración Propia

Por último, se arrastra el componente "**conexión_mysql_carga**", que contiene la conexión a la base de datos, que será donde se van a cargar los datos, y después se ejecuta TDI, como se ve en la figura 70.

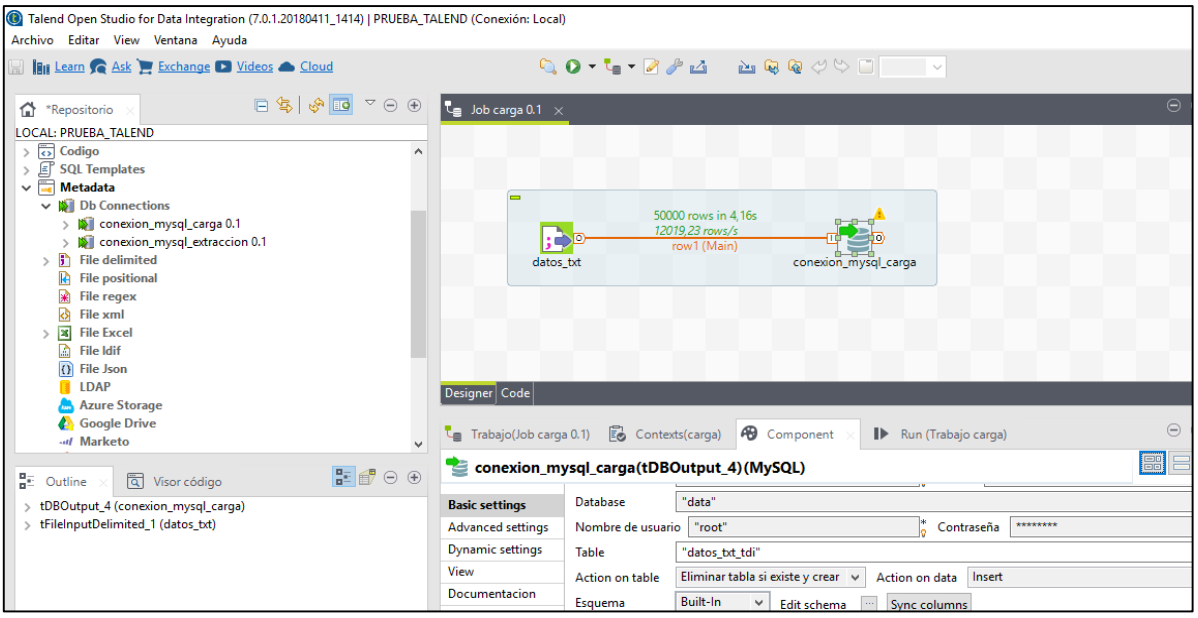


Figura 70: Carga de datos a una tabla de MySQL con TDI
Fuente: Elaboración Propia

En cuanto a velocidad de lectura y escritura de datos, tiempo que tarda en ejecución del proceso y facilidad de carga de datos, luego de ejecutar PDI, arroja como resultado que los datos se cargaron en un tiempo de 4.16 segundos a una velocidad de 12.019 registros por segundo. En cuando a la facilidad de carga de datos se puede decir que es el más complejo en la primera interacción con la herramienta, ya que se debe crear dos esquemas de metadatos uno que son los datos de entrada, en este caso los del archivo plano y uno que son los datos de salida, en este caso los de la tabla de MySQL, también se debe tener en cuenta que los tipos de datos por ejemplo (Fecha, entero, string, entre otros) concuerden en la base de datos para evitar cargar mal los datos o que no se carguen. Los datos se pueden ver en la figura 71.

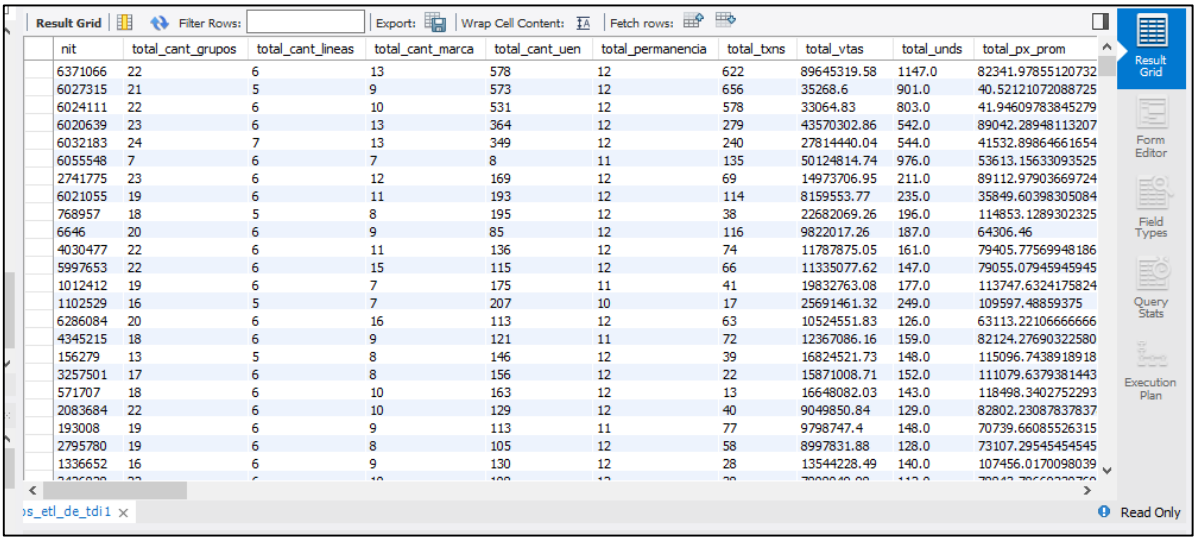


Figura 71: Datos cargados a la tabla de MySQL con TDI
Fuente: Elaboración Propia

En conclusión TDI, no solo permite cargar datos a una tabla de MySQL, sino también a muchos más Sistemas Gestores de Base de Datos como (Oracle, DB2, PostgreSQL,

Teradata, entre otros), permite cargar datos en archivos de Excel (xlsx, xls), archivos planos (csv, txt), archivos JSON, archivos XML, entre otros. Es un gran punto que se debe tener en cuenta a la hora de poder elegir una herramienta, con el proposito de realizar proceso ETL.

7.3 Carga de datos a una tabla de MySQL con OR

Para realizar en OR la carga de datos a una tabla de MySQL, primero se debe extraer los datos del archivo plano, para eso debe ir al componente "**Este equipo**", y dar clic en el botón "**Elegir archivos**", y luego buscar el archivo, como se ve en la figura 72.

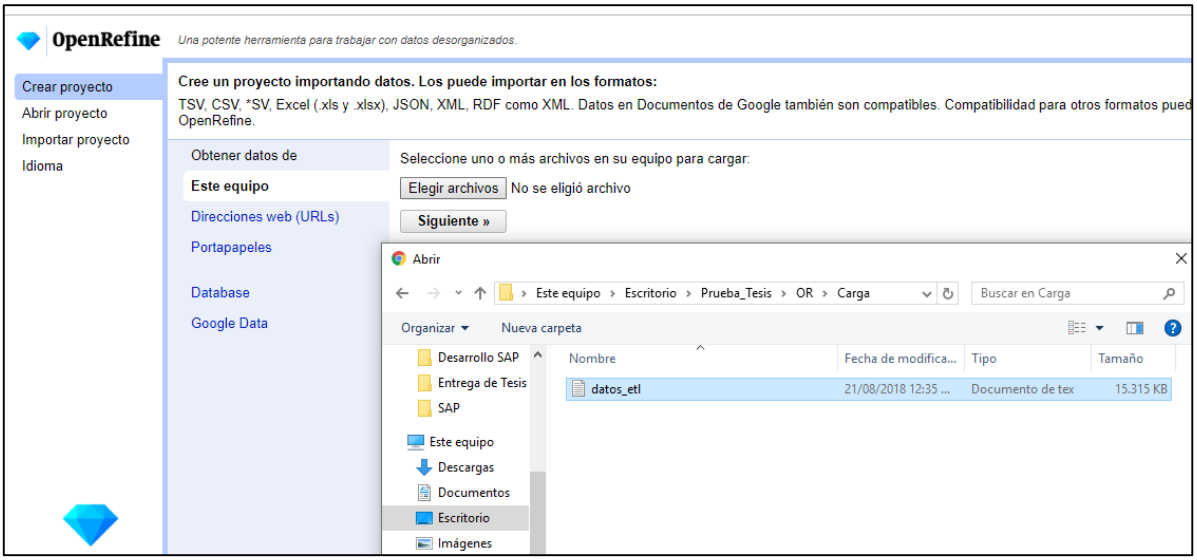


Figura 72: Extracción de los datos del archivo plano con OR
Fuente: Elaboración Propia

Después de extraer los datos, se crea un proyecto llamado carga, como se puede ver en la figura 73.

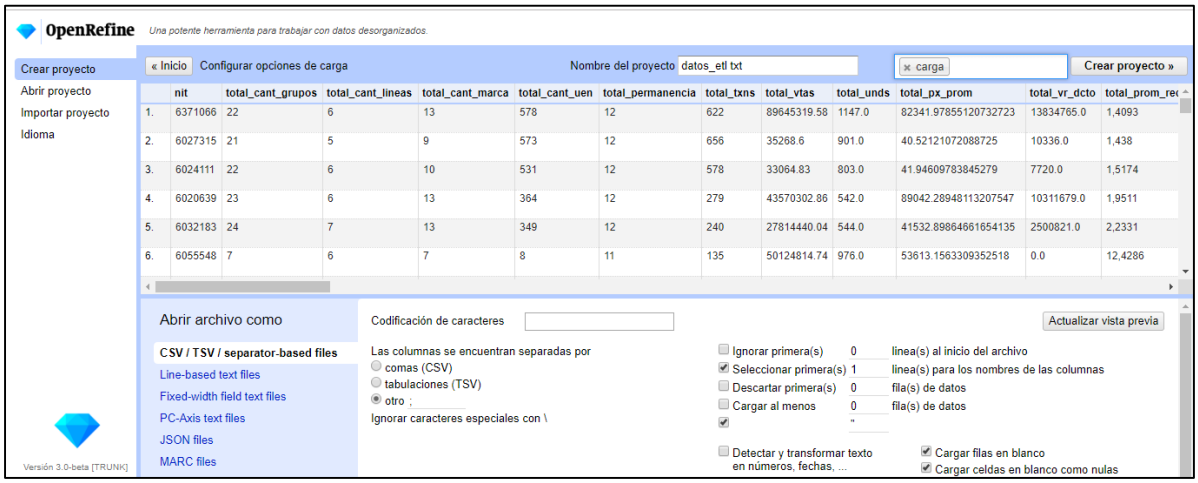


Figura 73: Creación del proyecto con OR
Fuente: Elaboración Propia

Por último, luego de haber extraído los datos, se prosigue a cargarlos en una tabla de la base de datos MySQL, para ello se va al componente "**Exportar**", y podemos observar que OR no tiene la capacidad para cargar los datos en una tabla de base de datos, como se puede ver

en la figura 74. Es decir, OR tiene una desventaja en cuanto a la carga de datos, porque solo permite cargar los datos en archivos planos, archivos de Excel, archivos XML, entre otros.

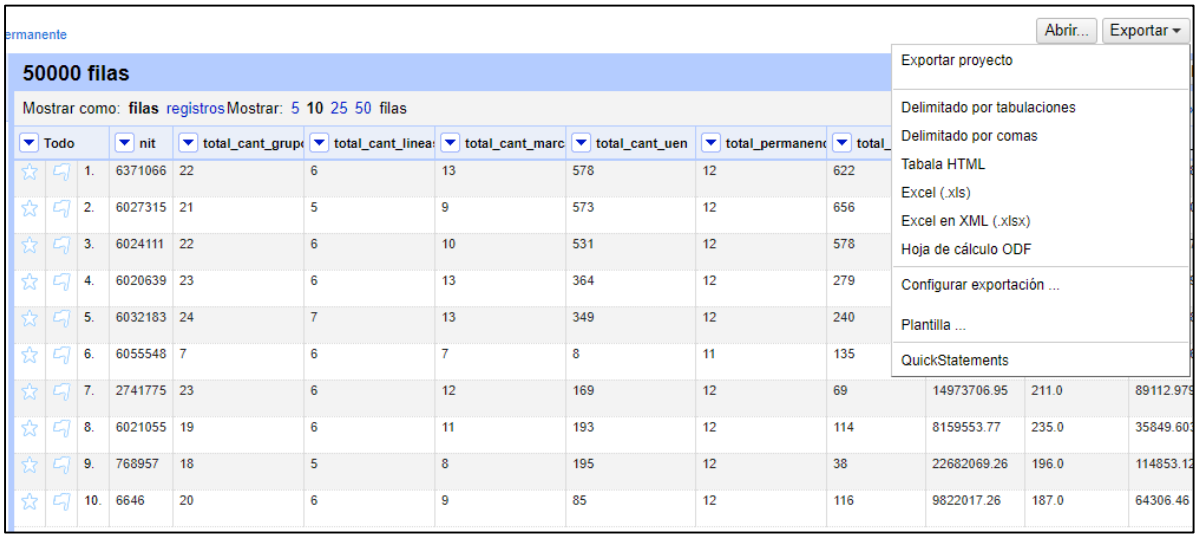


Figura 74: Carga de datos a una tabla de MySQL con OR
Fuente: Elaboración Propia

En cuanto a la velocidad de lectura y escritura de datos, tiempo que tarda en ejecución del proceso y facilidad de carga de datos, se puede decir que OR no maneja un registro histórico de lectura y escritura de datos, ni tampoco registra el tiempo que tarda en ejecución del proceso, pero en cuanto a la facilidad se puede decir, que es la herramienta más fácil de usar, porque es muy intuitiva y no se requiere un conocimiento profundo de bases de datos para poder utilizarla.

Para el caso de OR, no permite cargar los datos a una base de datos en la versión que tiene actualmente, como se ve en la figura 75, dándole una desventaja con respecto a las demás, es decir, en este caso solo transformaría los datos en esta herramienta, que luego se exporta en un archivo plano, que puede cargar con alguna de las dos herramientas antes mencionadas a una base de datos.

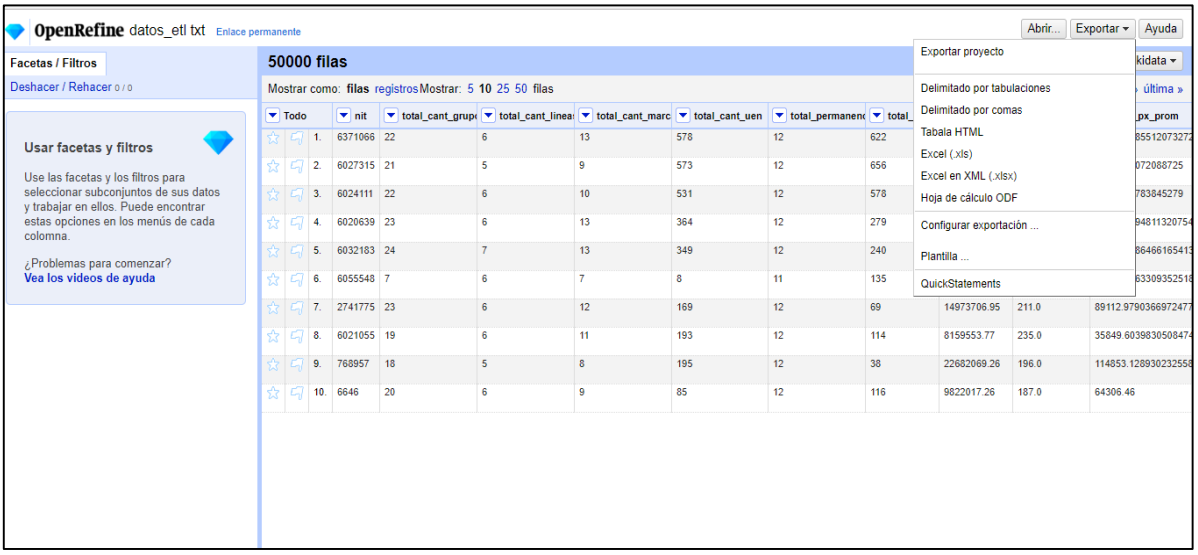


Figura 75: Fuentes a la que se puede cargar datos con OR
Fuente: Elaboración Propia

7.4 Carga de datos a un archivo de Excel

Para el caso en PDI, cargar los datos a un archivo de Excel, se debe primero leer los datos del archivo plano, donde luego se cargan a un archivo de Excel. Como se puede notar en la figura 76.

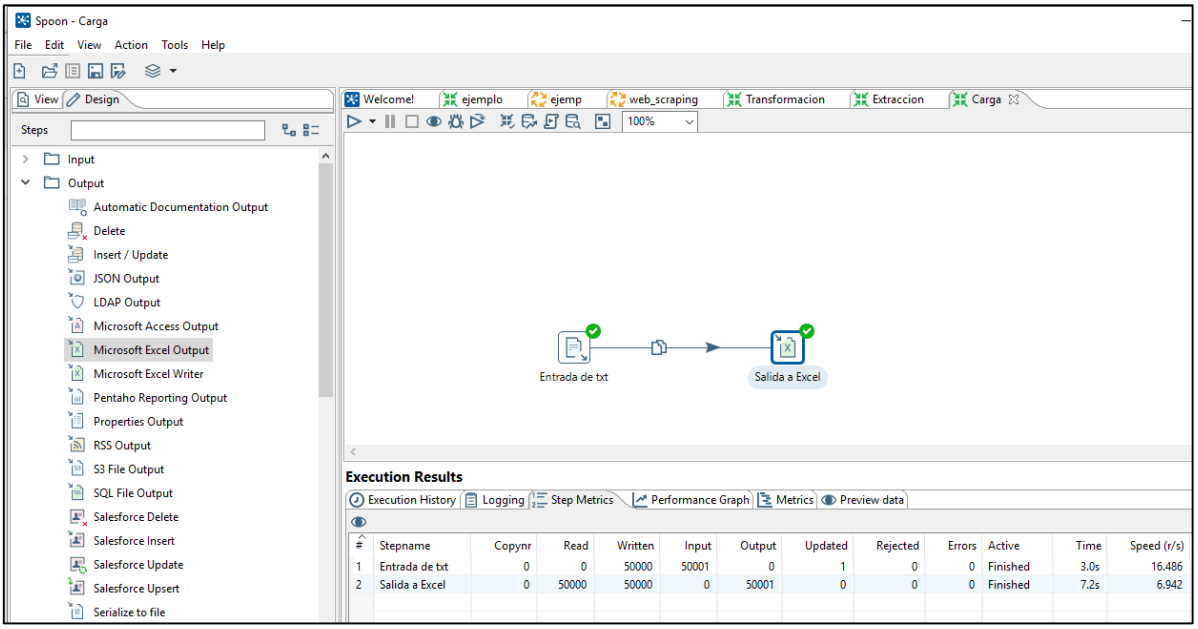


Figura 76: Carga de datos a un archivo de Excel con PDI
Fuente: Elaboración Propia

Los datos se extrajeron de un archivo plano en un tiempo de 3.0 segundos a una velocidad de 16.486 registros por segundo, que luego se cargaron en un archivo de Excel en un tiempo de 7.2 segundos a una velocidad de 6.942 registros por segundo. Los datos se pueden ver en la figura 77. En cuando a la facilidad de carga de datos se puede decir que es fácil en la primera interacción con la herramienta, es decir no requiere un conocimiento previo, porque solo es extraer los datos del archivo plano de origen y cargarlos a el archivo de destino Excel.

	A	B	C	D	E	F	G	H	I	J	K
1	nit	total_cant_grupos	total_cant_lineas	total_cant_marca	total_cant_uen	total_permanencia	total_txns	total_vtas	total_unds	total_px_prom	total_vr_dcto
2	6371066	22	6	13	578	12	622	89645319.58	1147.	82341.97855	13834765.
3	6027315	21	5	9	573	12	656	35268.6	901.	40.52121	10336.
4	6024111	22	6	10	531	12	578	33064.83	803.	41.9461	7720.
5	6020639	23	6	13	364	12	279	43570302.86	542.	89042.28948	10311679.
6	6032183	24	7	13	349	12	240	27814440.04	544.	41532.89865	2500821.
7	6055548	7	6	7	8	11	135	50124814.74	976.	53613.15633	0.
8	2741775	23	6	12	169	12	69	14973706.95	211.	89112.97904	4960570.
9	6021055	19	6	11	193	12	114	8159553.77	235.	35849.60398	1240530.
10	768957	18	5	8	195	12	38	22682069.26	196.	114853.12893	1539140.
11	6646	20	6	9	85	12	116	9822017.26	187.	64306.46	1574760.
12	4030477	22	6	11	136	12	74	11787875.05	161.	79405.7757	5600365.
13	5997653	22	6	15	115	12	66	11335077.62	147.	79055.07946	4223610.
14	1012412	19	6	7	175	11	41	19832763.08	177.	113747.63242	3408995.
15	1102529	16	5	7	207	10	17	25691461.32	249.	109597.48859	4825815.
16	6286084	20	6	16	113	12	63	10524551.83	126.	63113.22107	2918020.
17	4345215	18	6	9	121	11	72	12367086.16	159.	82124.2769	5555580.
18	156279	13	5	8	146	12	39	16824521.73	148.	115096.74389	2495720.
19	3257501	17	6	8	156	12	22	15871008.71	152.	111079.63794	2784430.
20	571707	18	6	10	163	12	13	16648082.03	143.	118498.34028	1109665.
21	2083684	22	6	10	129	12	40	9049850.84	129.	82802.23088	2212065.

Figura 77: Datos cargados a un archivo de Excel con PDI
Fuente: Elaboración Propia

Para el caso en TDI, cargar los datos a un archivo de Excel, se debe primero crear un esquema de metadatos que es donde se van a extraer los datos, y luego se prosigue a crear otro esquema de destino que es donde se van a cargar los datos.

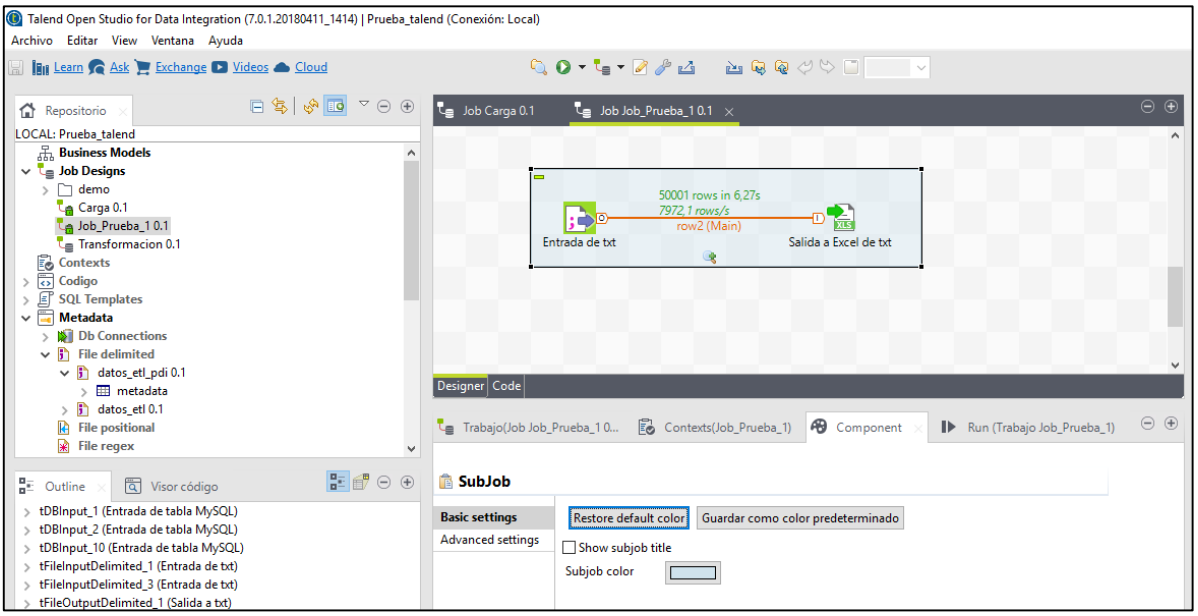
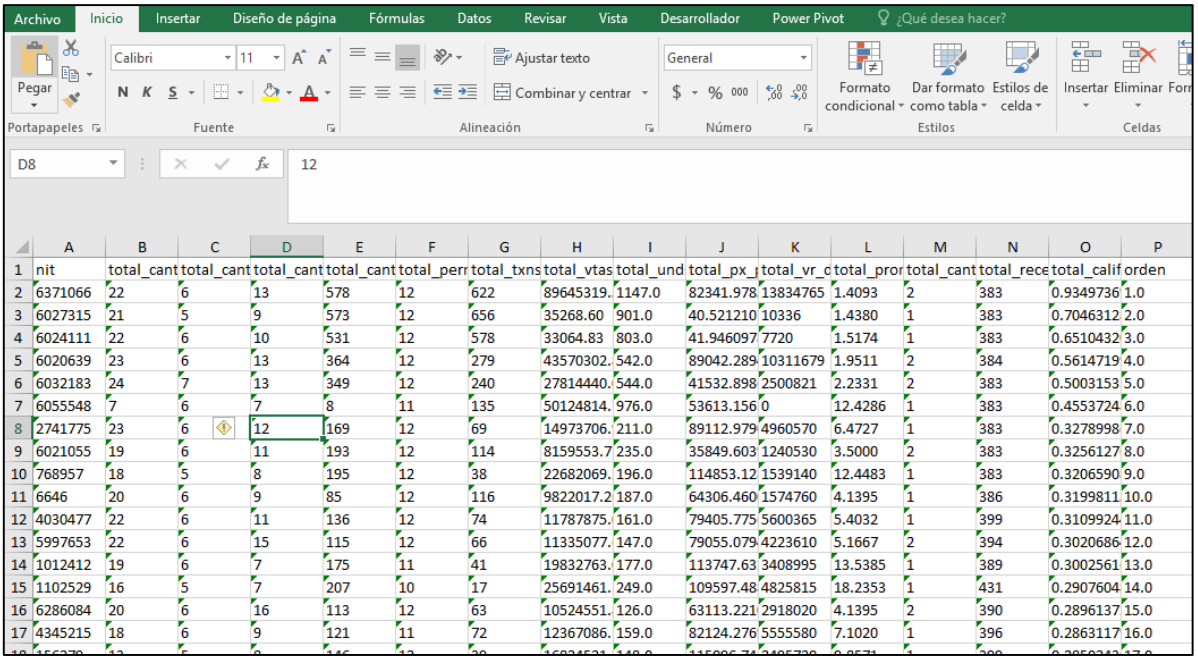


Figura 78: Carga de datos a un archivo de Excel con TDI
Fuente: Elaboración Propia

Como se puede notar en la figura 78, los datos se cargaron a un archivo de Excel, en un tiempo de 6.27 segundos a una velocidad de 7.972 registros por segundo. Los datos se pueden ver en la figura 79. En cuando a la facilidad de carga de datos se puede decir que tiene un grado de complejidad más a la hora de realizar el proceso que otra herramienta como PDI en la primera interacción, es decir requiere un conocimiento previo de los datos, para su posterior creación del esquema de metadatos.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	nit	total_cant	total_cant	total_cant	total_cant	total_perr	total_txns	total_vtas	total_und	total_px_j	total_vr_d	total_pror	total_cant	total_rece	total_calif	orden
2	6371066	22	6	13	578	12	622	89645319.	1147.0	82341.978	13834765	1.4093	2	383	0.9349736	1.0
3	6027315	21	5	9	573	12	656	35268.60	901.0	40.521210	10336	1.4380	1	383	0.7046312	2.0
4	6024111	22	6	10	531	12	578	33064.83	803.0	41.946097	7720	1.5174	1	383	0.6510432	3.0
5	6020639	23	6	13	364	12	279	43570302.	542.0	89042.289	10311679	1.9511	2	384	0.5614719	4.0
6	6032183	24	7	13	349	12	240	27814440.	544.0	41532.898	2500821	2.2331	2	383	0.5003153	5.0
7	6055548	7	6	7	8	11	135	50124814.	976.0	53613.156	0	12.4286	1	383	0.4553724	6.0
8	2741775	23	6	12	169	12	69	14973706.	211.0	89112.979	4960570	6.4727	1	383	0.3278998	7.0
9	6021055	19	6	11	193	12	114	8159553.7	235.0	35849.603	1240530	3.5000	2	383	0.3256127	8.0
10	768957	18	5	8	195	12	38	22682069.	196.0	114853.12	1539140	12.4483	1	383	0.3206590	9.0
11	6646	20	6	9	85	12	116	9822017.2	187.0	64306.460	1574760	4.1395	1	386	0.3199811	10.0
12	4030477	22	6	11	136	12	74	11787875.	161.0	79405.775	5600365	5.4032	1	399	0.3109924	11.0
13	5997653	22	6	15	115	12	66	11335077.	147.0	79055.079	4223610	5.1667	2	394	0.3020686	12.0
14	1012412	19	6	7	175	11	41	19832763.	177.0	113747.63	3408995	13.5385	1	389	0.3002561	13.0
15	1102529	16	5	7	207	10	17	25691461.	249.0	109597.48	4825815	18.2353	1	431	0.2907604	14.0
16	6286084	20	6	16	113	12	63	10524551.	126.0	63113.221	2918020	4.1395	2	390	0.2896137	15.0
17	4345215	18	6	9	121	11	72	12367086.	159.0	82124.276	5555580	7.1020	1	396	0.2863117	16.0

Figura 79: Datos cargados a un archivo de Excel con TDI
Fuente: Elaboración Propia

Para el caso de OR, cargar los datos a un archivo de Excel, se debe primero extraer los datos del archivo plano, luego crear un proyecto que permita cargar los datos a un archivo de Excel, como se puede ver en la figura 80.

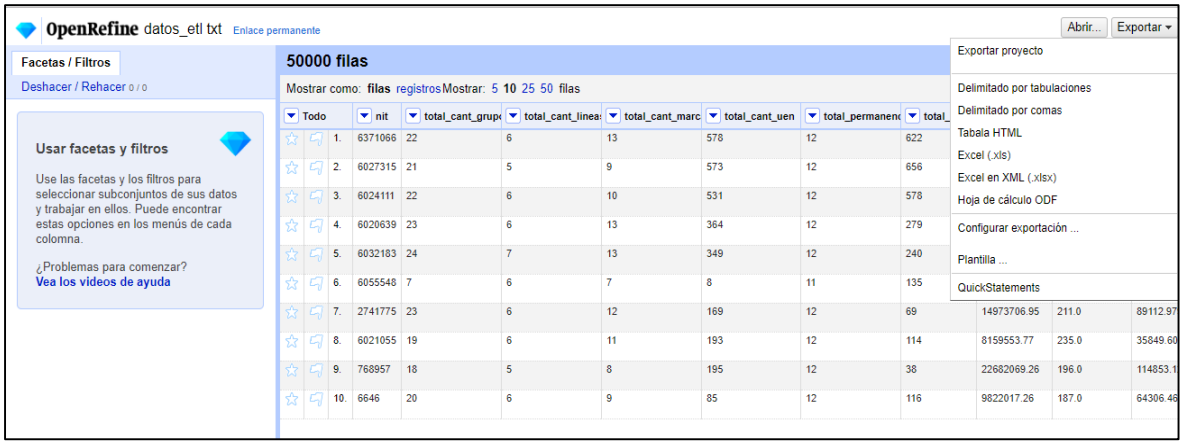
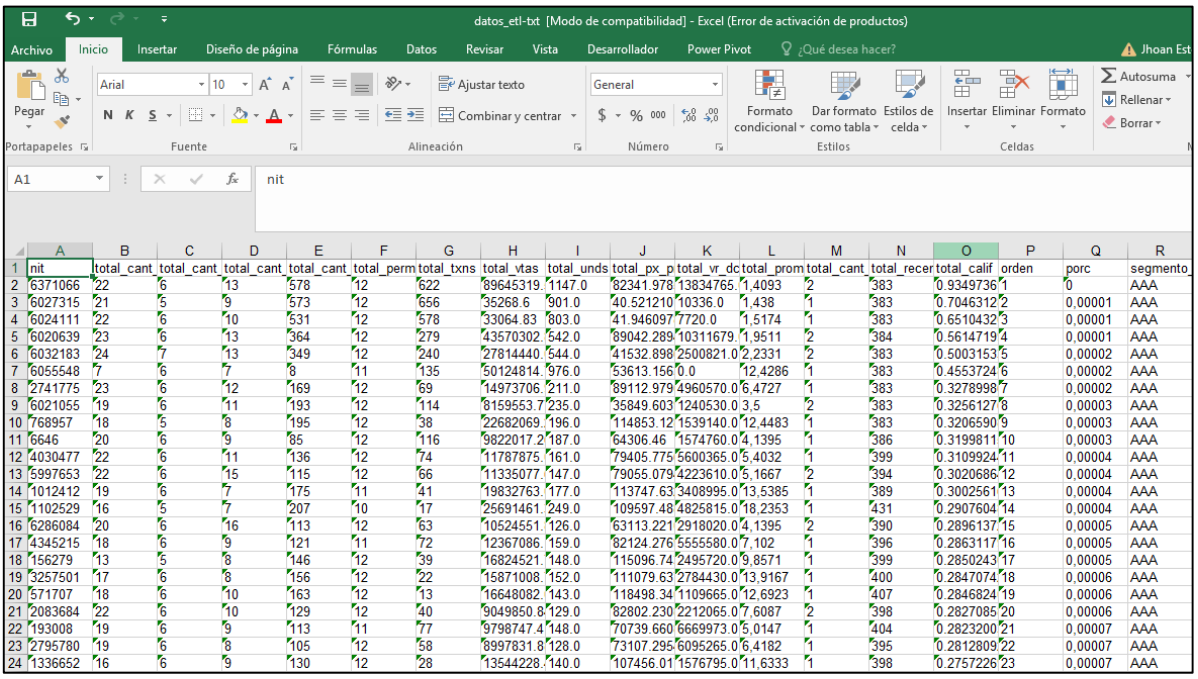


Figura 80: Carga de datos a un archivo de Excel con OR
Fuente: Elaboración Propia

Como se puede notar en la figura 81, los datos se cargaron a un archivo de Excel, lo que no permite OR es medir el tiempo de ejecución ni la velocidad de carga de datos como las demás herramientas, pero se pudo cargar los datos a un archivo de Excel con extensión (xls) y no se pudo cargar los datos para la extensión (xlsx) ya que el servidor no soporto la carga de datos.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	nit	total_cant	total_cant	total_cant	total_cant	total_perm	total_txns	total_vtas	total_unds	total_px_p	total_vr_dc	total_prom	total_cant	total_recer	total_calif	orden	porc	segmento_
2	6371066	22	6	13	578	12	622	89645319.	1147.0	82341.978	13834765.	1,4093	2	383	0.9349736	1	0	AAA
3	6027315	21	5	9	573	12	656	35268.6	901.0	40.521210	10336.0	1,438	1	383	0.7046312	2	0,00001	AAA
4	6024111	22	6	10	531	12	578	33064.83	803.0	41.946097	7720.0	1,5174	1	383	0.6510432	3	0,00001	AAA
5	6020639	23	6	13	364	12	279	43570302.	542.0	89042.289	10311679.	1,9511	2	384	0.5614719	4	0,00001	AAA
6	6032183	24	7	13	349	12	240	27814440.	544.0	41532.898	2500821.0	2,2331	2	383	0.5003153	5	0,00002	AAA
7	6055548	7	6	7	8	11	135	50124814.	976.0	53613.156	0.0	12,4286	1	383	0.4553724	6	0,00002	AAA
8	2741775	23	6	12	169	12	69	14973706.	211.0	89112.979	4960570.0	6,4727	1	383	0.3278998	7	0,00002	AAA
9	6021055	19	6	11	193	12	114	8159553.7	235.0	35849.603	1240530.0	3,5	2	383	0.3256127	8	0,00003	AAA
10	768957	18	5	8	195	12	38	22682069.	196.0	114853.12	1539140.0	12,4483	1	383	0.3206590	9	0,00003	AAA
11	6646	20	6	9	85	12	116	9822017.2	187.0	64306.46	1574760.0	4,1395	1	386	0.3199811	10	0,00003	AAA
12	4030477	22	6	11	136	12	74	11787875.	161.0	79405.775	5600365.0	5,4032	1	399	0.3109924	11	0,00004	AAA
13	5997653	22	6	15	115	12	66	11335077.	147.0	79055.079	4223610.0	5,1667	2	394	0.3020686	12	0,00004	AAA
14	1012412	19	6	7	175	11	41	19832763.	177.0	113747.63	3408995.0	13,5385	1	389	0.3002561	13	0,00004	AAA
15	1102529	16	5	7	207	10	17	25691461.	249.0	109597.48	4825815.0	18,2353	1	431	0.2907604	14	0,00004	AAA
16	6286084	20	6	16	113	12	63	10524551.	126.0	63113.221	2918020.0	4,1395	2	390	0.2896137	15	0,00005	AAA
17	4345215	18	6	9	121	11	72	12367086.	159.0	82124.276	5555580.0	7,102	1	396	0.2863117	16	0,00005	AAA
18	156279	13	5	8	146	12	39	16824521.	148.0	115096.74	2495720.0	9,8571	1	399	0.2850243	17	0,00005	AAA
19	3257501	17	6	8	156	12	22	15871008.	152.0	111079.63	2784430.0	13,9167	1	400	0.2847074	18	0,00006	AAA
20	571707	18	6	10	163	12	13	16648082.	143.0	118498.34	1109665.0	12,6923	1	407	0.2846824	19	0,00006	AAA
21	2083684	22	6	10	129	12	40	9049850.8	129.0	82802.230	2212065.0	7,6087	2	398	0.2827085	20	0,00006	AAA
22	193008	19	6	9	113	11	77	9798747.4	148.0	70739.660	6669973.0	5,0147	1	404	0.2823200	21	0,00007	AAA
23	2795780	19	6	8	105	12	58	8997831.8	128.0	73107.295	6095265.0	6,4182	1	395	0.2812809	22	0,00007	AAA
24	1336652	16	6	9	130	12	28	13544228.	140.0	107456.01	1576795.0	11,6333	1	398	0.2757226	23	0,00007	AAA

Figura 81: Datos cargados a un archivo de Excel con OR
Fuente: Elaboración Propia

La figura 81, presenta los datos que fueron cargados de forma adecuada, pero en extensión (xls). En cuanto a las figuras 82 y 83 se puede observar que OR, requiere de buenos recursos como procesador y RAM, cuando hablamos de cargar gran cantidad de datos, en este caso se habla de 50.000 registros y 61 campos, que es la base usada para ilustrar.

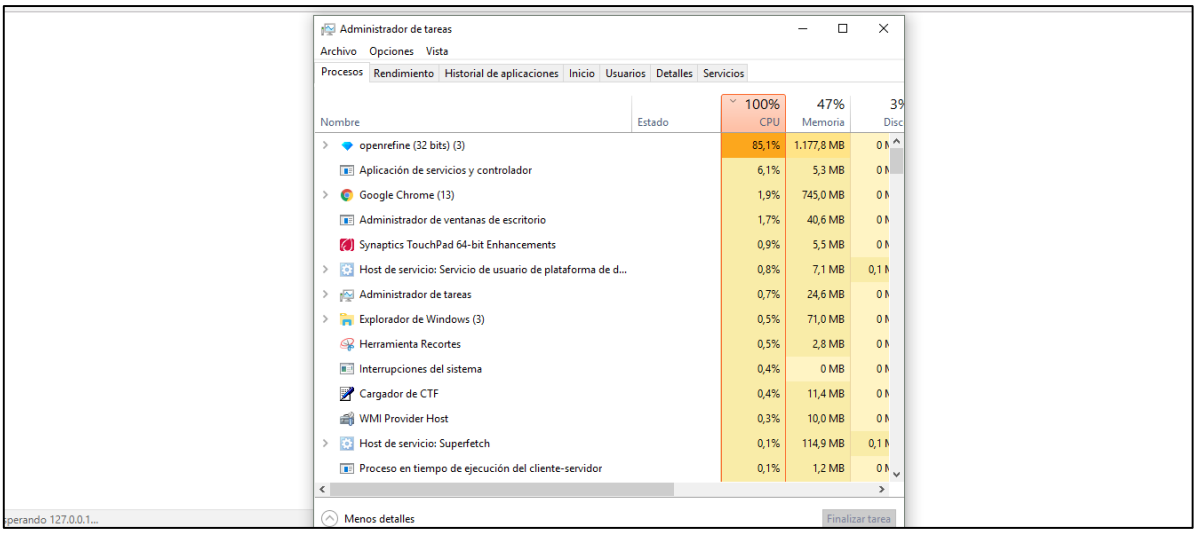


Figura 82: Consumo de CPU y RAM en la carga de datos a un archivo de Excel con OR
Fuente: Elaboración Propia

HTTP ERROR 500

Problem accessing /command/core/export-rows/datos_etl-txt.xlsx. Reason:

```
GC overhead limit exceeded
```

Caused by:

```
java.lang.OutOfMemoryError: GC overhead limit exceeded
        at org.apache.xmlbeans.impl.store.Saver$TextSaver.resize(Saver.java:1592)
        at org.apache.xmlbeans.impl.store.Saver$TextSaver.preEmit(Saver.java:1223)
        at org.apache.xmlbeans.impl.store.Saver$TextSaver.emit(Saver.java:1144)
        at org.apache.xmlbeans.impl.store.Saver$TextSaver.emitElement(Saver.java:926)
        at org.apache.xmlbeans.impl.store.Saver.processElement(Saver.java:456)
        at org.apache.xmlbeans.impl.store.Saver.process(Saver.java:307)
        at org.apache.xmlbeans.impl.store.Saver$TextSaver.saveToString(Saver.java:1727)
        at org.apache.xmlbeans.impl.store.Cursor._xmlText(Cursor.java:546)
        at org.apache.xmlbeans.impl.store.Cursor.xmlText(Cursor.java:2436)
        at org.apache.xmlbeans.impl.values.XmlObjectBase.xmlText(XmlObjectBase.java:1455)
        at org.apache.poi.xssf.model.SharedStringsTable.getKey(SharedStringsTable.java:134)
        at org.apache.poi.xssf.model.SharedStringsTable.addEntry(SharedStringsTable.java:180)
        at org.apache.poi.xssf.usermodel.XSSFCell.setCellValue(XSSFCell.java:369)
        at org.apache.poi.xssf.usermodel.XSSFCell.setCellValue(XSSFCell.java:333)
        at com.google.refine.exporters.XlsExporter$1.addRow(XlsExporter.java:125)
        at com.google.refine.exporters.CustomizableTabularExporterUtilities$1.visit(CustomizableTabularExporterUtilities.java:157)
        at com.google.refine.browsing.util.ConjunctiveFilteredRows.visitRow(ConjunctiveFilteredRows.java:76)
        at com.google.refine.browsing.util.ConjunctiveFilteredRows.accept(ConjunctiveFilteredRows.java:65)
        at com.google.refine.exporters.CustomizableTabularExporterUtilities.exportRows(CustomizableTabularExporterUtilities.java:171)
        at com.google.refine.exporters.XlsExporter.export(XlsExporter.java:139)
        at com.google.refine.commands.project.ExportRowsCommand.doPost(ExportRowsCommand.java:116)
        at com.google.refine.RefineServlet.service(RefineServlet.java:178)
        at javax.servlet.http.HttpServlet.service(HttpServlet.java:820)
        at org.mortbay.jetty.servlet.ServletHolder.handle(ServletHolder.java:511)
        at org.mortbay.jetty.servlet.ServletHandler$CachedChain.doFilter(ServletHandler.java:1166)
        at org.mortbay.servlet.UserAgentFilter.doFilter(UserAgentFilter.java:81)
        at org.mortbay.servlet.GzipFilter.doFilter(GzipFilter.java:132)
```

Figura 83: Sistema caído en la carga de datos a un archivo de Excel con OR
Fuente: Elaboración Propia

7.5 Conclusión de carga de datos con PDI, TDI y OR

En conclusión, PDI fue la herramienta menos compleja de utilizar, sin embargo, TDI fue la herramienta más eficiente en cuando al tiempo de ejecución, velocidad de lectura y escritura de datos, siendo está un poco más compleja de utilizar a la hora de cargar archivos planos o archivos de Excel. Se aclara que en caso que OR pudiera cargar los datos a base de datos y el servidor soportará la carga a archivo de Excel en formato (xlsx), está sería la herramienta más fácil y sencilla de utilizar. En la tabla 7.1 se compara las herramientas con algunos componentes que los hace diferentes en cuanto a la carga de datos.

Herramienta / Proceso	Talend Data Integration (TDI)	Pentaho Data Integration (PDI)	OpenRefine (OR)
Facilidad de carga datos	Facilidad baja	Facilidad media	Facilidad alta
Permite la cargar de datos a bases de datos	Si	Si	No
Tiempo de ejecución en la carga del archivo plano a una tabla de MySQL (50.000 registros y 61 campos)	Rápido (4.16 Segundos)	Rápido (18.0 Segundos)	No tiene capacidad de medir el tiempo
Velocidad de carga del archivo plano a una tabla de MySQL (50.000 registros y 61 campos)	12.019 registros por segundo	2.780 registros por segundo	No tiene capacidad de medir la velocidad
Tiempo de ejecución en la carga del archivo plano a un archivo de Excel (50.000 registros y 61 campos)	Rápido (6.27 Segundos)	Rápido (7.2 Segundos)	No tiene capacidad de medir el tiempo
Velocidad de carga del archivo plano a un archivo de Excel (50.000 registros y 61 campos)	7.972 registros por segundo	6.942 registros por segundo	No tiene capacidad de medir la velocidad
Necesita generar un esquema de metadatos para la carga de datos	Si (Metadatos)	No	No
Carga de datos a una tabla de MySQL	Si	Si	No
Hoja de cálculo ODF	No	No	Si
Archivo ARFF	Si	No	No
Salesforce	Si	Si	No
Nube	**Carga de datos a la Nube**		
Azure Storage	Si	No	No
Google Storage	Si	No	No
Amazon	Si	No	No
Big Data	**Carga de datos a Big Data**		
Hive	Si	No	No
SSTable	No	Si	No
Cassandra	No	Si	No
MongoDB	No	Si	No
HBase	No	Si	No

Tabla 7-1: Comparación de las herramientas en la componente de Carga de datos usando la base de datos de retail del archivo plano.

En la tabla 7.1, se puede notar que PDI y TDI, no solo cargan datos a una base de datos o un archivo de Excel, sino que también pueden cargar datos a muchas fuentes, como bases de datos no relacionales, archivos XML, archivos de Excel, JSON, la nube entre otros. Mientras que OR es más limitado a la hora de cargar datos ya que solo puede cargar datos a algunos archivos planos.

8 Comparación de herramientas ETL, en otros aspectos

A continuación, se describirán más capacidades, que hacen que las herramientas tengan ventajas y desventajas adicionales sobre las otras.

8.1 Comparación de PDI y TDI con respecto a OR

PDI y TDI tiene una ventaja con respecto a OR, y es que permiten instalar complementos que desarrolla la comunidad, en PDI están disponibles los scripts de R y Python que le permite interactuar con ambos lenguajes de programación y complementar el proceso con los análisis que se pueden obtener de cualquiera de estas dos herramientas, por otro lado TDI permite una comunicación con SPSS de IBM es decir que también puede aumentar su capacidad de análisis a la hora de necesitarse, en cambio OR tiene como potencialidad poder transformar variables por medio de Python/Jypython. También cabe mencionar que como TDI, PDI y OR están hechos en Java, se pueden crear rutinas que potencialicen las capacidades de ambas herramientas dando mayor versatilidad a la hora de realizar algún proceso que no esté contemplado en las funcionalidades.

PDI y TDI con respecto a OR, poseen una mejor conexión con herramientas de Big Data, teniendo en cuenta que TDI está muy relacionado con Google y Hadoop en Big Data (BigQuery, Storage y Hive), mientras que PDI está muy relacionado con Hadoop, MapReduce y las bases de datos NoSQL (Avro, Cassandra, CouchDb, HBase, MongDB y SSTable). Además, TDI posee la capacidad de procesar y conectarse a la nube, hecho que también poseen PDI, pero no con igual potencialidad, ya que TDI, permite conectarse a Amazon, Azure Storage, Dropbox, Google Storage, entre otras.

PDI y TDI con respecto a OR son herramientas más orientadas al proceso ETL completo, mientras que OR es una herramienta más orientada la transformación de datos, es decir en cuanto a la extracción y carga de datos, tiene mucha desventaja con respecto a PDI y TDI, que poseen grandes capacidades en la extracción, transformación y carga de datos.

8.2 Ventajas y desventajas de PDI, TDI y OR

Una diferencia entre todas las herramientas es que PDI trabaja básicamente con dos componentes como lo son trabajos (**Job)** y transformaciones (**Transfomation**), en cambio TDI solo tiene la componente de trabajo (**Job**), mientras que OR posee una estructura de proyectos en su metodología de trabajo, que le permite guardar los pasos e incluso reservar acciones.

Una ventaja que posee TDI con respecto a las demás herramientas es que posee dos componentes diferenciadores como los son (Business Intelligence y Business) que le permiten realizar procesos relacionados con el negocio dándole una versatilidad a la hora de exponer cualquiera de estas funcionalidades que las demás herramientas no poseen. Pero una de las ventajas que tiene OR con respecto a las demás herramientas y que es de gran importancia es la forma en que permite procesar los datos, de una forma visual, fácil y sencilla como Excel.

Una ventaja que tiene TDI sobre las demás herramientas es que tiene unas funcionalidades llamadas *Business Models* y *Documentation*, que permiten crear esquemas a modo de documentación del proyecto, es decir se pueden describir todos los procesos que se están

desarrollando en un determinado proyecto y a su vez se puede ir consultando la información del estado del proyecto. TDI también posee una funcionalidad llamada contexto que es un conjunto de variables que se utiliza para determinar el comportamiento de un Job, es comúnmente usado para diferenciar entre entornos como el desarrollo, preproducción y producción.

TDI es una herramienta un poco más compleja de aprender que las demás, teniendo en cuenta que para procesar datos siempre se necesita generar un esquema de Metadatos, que le permita realizar el proceso que se requiere, que a su vez es una ventaja de tener los esquemas de metadatos organizados.

A continuación, se hace referencia a dos tablas que comparan las capacidades y características relevantes de las herramientas.

Herramienta / Proceso	Talend Data Integration (TDI)	Pentaho Data Integration (PDI)	OpenRefine (OR)
Script de R	NO	SI	NO
Script de Python	NO	SI	NO
Big Data	SI (Limitado a herramientas de Google y Hive)	SI (Tiene muchas más herramientas de soporte)	NO
Conexión a SPSS	SI	NO	NO
Business Models	SI	NO	NO
Documentation	SI	NO	NO

Tabla 8-1: Tabla de comparación de herramientas ETL en capacidades

Herramienta / Proceso	Talend Data Integration (TDI)	Pentaho Data Integration (PDI)	OpenRefine (OR)
Costo	OpenSource (Tiene versión de pago)	OpenSource (Tiene versión de pago)	OpenSource
Riesgo	Alto	Medio	Bajo
Facilidad de uso	Facilidad baja	Facilidad media	Facilidad alta
Visualización	Baja	Media	Alta
Conectividad	Alta	Alta	Baja
Velocidad de lectura y escritura de datos	Alta	Alta	Baja
Tiempo que tarda en ejecución del proceso	Alta	Alta	Baja
Soporte	SI (En versión de pago)	SI (En versión de pago)	NO
Implementación	Media	Media	Alta
Monitoreo	Alta	Alta	Bajo

Tabla 8-2: Comparación de herramientas ETL en características

Costo de la herramienta

- Significa el costo promedio de cierto producto. Desde costo de orden, licencia, servicio, soporte, entrenamiento, consultoría y cualquier otro pago adicional, que se tenga que realizar para el uso total de la herramienta.
- Las herramientas de código abierto son naturalmente gratis de utilizar, pero el soporte, entrenamiento y consultoría son los costos a considerar.

Costo

En este caso TDI y PDI son softwares comerciales, y cuentan con versiones de código abierto, como también es el caso de OR, donde solo hay versión de código abierto.

Riesgo

Cuando hablamos de herramientas de código abierto siempre hay un riesgo cuando se habla de manipulación de datos, para este caso hablamos de:

- Falta capacitación para uso de herramienta
- No cumplir con requerimiento o expectativas
- Soporte de la herramienta

Facilidad de uso

- TDI cuenta con una interfaz de usuario (GUI) que se base en un add-on para Eclipse, que puede ser intuitiva, pero una persona nueva en la utilización de la herramienta siempre tendrá un grado de dificultad mientras se adapta a su manejo.
- PDI cuenta con una interfaz de usuario (GUI) que es sencilla de utilizar, pero una persona nueva en la utilización de la herramienta siempre tendrá un grado de dificulta mientras se adapta a su manejo.
- OR cuenta con la interfaz de usuario (GUI) que es sencilla e intuitiva para utilizar, aunque sus capacidades son limitadas por ese motivo el usuario decide que herramienta utilizar dependiendo de lo que busca solucionar.

Visualización

- OR es la herramienta que permite visualizar los datos de una forma más sencilla y fácil como si fuera tipo Excel.
- PDI permite visualizar de una forma sencilla, pero no es igual de fácil como en OR.
- TDI es la herramienta que posee más complejidad para visualizar los datos, porque se visualizan por consola con una funcionalidad.

Conectividad

- TDI posee conectividad a varias bases de datos, archivos planos, xml, Excel, servicios web, la nube, big data, entre otros.
- PDI posee conectividad a varias bases de datos, archivos planos, xml, Excel, servicios web, la nube, big data, entre otros.
- OR posee conectividad a varias bases de datos, archivos planos, xml, Excel, entre otros, siendo de las tres la herramienta más limitada en conexiones.

Velocidad de lectura y escritura de datos

- PDI es más rápido que TDI en el proceso de Extracción de datos (E), mientras que TDI es más rápido en el proceso de Transformación y Carga de datos (L), además TDI requiere configuración específica y manual, con conocimiento previo de los datos a utilizar.
- Como OR no depende de pasos, en este caso acelera el proceso ETL, pero en conexión con bases de datos a la hora de consultar una base de datos con muchos registros se demora un tiempo considerablemente alto y hay momentos donde se cae el servidor.
- En general se puede decir que TDI es más rápido que PDI en el proceso ETL.

Tiempo que tarda en ejecución del proceso

- TDI es la herramienta que posee un mejor rendimiento en el tiempo de ejecución, en cuanto el proceso de Transformación (T) y Carga de datos (L).
- PDI también tiene un buen rendimiento en el tiempo de ejecución en el proceso Extracción (E).
- OR posee un rendimiento adecuado en comparación con las demás herramientas, siendo el de menor rendimiento, en tiempo de ejecución.
- En general se puede decir que TDI tiene mejor rendimiento en el tiempo de ejecución del proceso ETL.

Soporte

- TDI cuenta con soporte, pero en este caso como se comparan es las herramientas de código abierto, en este caso no posee soporte, y si desea comprar la licencia en ese caso se podrá disponer del soporte adecuado.
- PDI cuenta con soporte, pero en este caso como se comparan es las herramientas de código abierto, en este caso no posee soporte, y si desea comprar la licencia en ese caso se podrá disponer del soporte adecuado.
- OR no cuenta con soporte.

Implementación

Herramienta / Proceso	Talend Data Integration (TDI)	Pentaho Data Integration (PDI)	OpenRefine (OR)
PLATAFORMA	Cualquiera compatible con Java o Perl (Windows, Linux, Mac)	Cualquiera compatible con Java (Windows, Linux, Mac)	(Windows, Linux, Mac)
RAM	512 MB	512 MB	512 MB
CPU	1.2 GHZ	1.2 GHZ	1 GHZ
Lenguajes	Java	Java	Java

Tabla 8-3: Características de implementación de herramientas ETL

En la implementación cabe aclarar, que OR es una herramienta que requiere tener muy buenas capacidades del PC tanto en procesador, mínimo Core i7, como en RAM, mínimo 8gb, para un mejor rendimiento.

Monitoreo

- TDI tiene herramientas prácticas de monitoreo y registro histórico
- PDI tiene herramientas prácticas de monitoreo y registro histórico
- OR no tiene herramientas prácticas de monitoreo y registro histórico, pero si posee un registro de pasos que se va efectuando al hacer cualquier manipulación.

9 Recomendaciones para seleccionar la herramienta más adecuada de las comparadas, que permita más beneficios a la hora de utilizarla según sea el caso Extraer, Transformar o Cargar datos.

Con base en la comparación en la comparación de las herramientas de ETL, se hace unas sugerencias de los casos en que puede elegir una u otra herramienta:

Extracción de datos (Data Extraction)

- Si un usuario busca una herramienta que le permita realizar el proceso de extracción de datos, en este caso por la característica de la herramienta, si desea extraer datos de Bases de Datos, Archivos Planos, Archivos de Excel, entre otros se puede dirigir a Pentaho Data Integration, por la comodidad y facilidad que maneja para la extracción de datos de todas estas fuentes.
- Si un usuario busca la Extracción de datos de la nube, por las características de la herramienta se debe dirigir Talend Data Integration, aunque también sirve para extraer datos de Bases de Datos, Archivos Planos, Archivos de Excel, entre otros, pero tiene un grado más de complejidad que PDI.
- Si un usuario busca la Extracción de datos, por la característica de la herramienta OpenRefine es la menos indicada para esta labor, pero es la más fácil y sencilla de usar, sin importar que en algunos casos pueden ser limitada en las fuentes de donde se pueden extraer datos, ya que posee muy pocas con respecto a las demás herramientas.
- Si un usuario busca la Extracción de datos, en este caso por la característica de la herramienta si desea extraer datos de Big Data se puede dirigir a Pentaho Data Integration, por la comodidad y facilidad que maneja para la extracción de datos de todas de estas fuentes, además te ofrece muchas más componentes en este campo que Talend Data Integration y OpenRefine.
- En conclusión, el usuario puede usar siempre la herramienta que desee y la que más le guste o este enseñado a usar, pero desde lo investigado no hay una recomendación única para el proceso de Extracción de datos, todo depende de lo que el usuario este buscando solucionar, en este caso solo se está sugiriendo que puede utilizar en un caso dado.

Transformación de datos (Data Transformation)

- Si un usuario busca la Transformación de datos, en este caso por la característica de la herramienta si desea transformar datos se puede dirigir a OpenRefine, por la comodidad, facilidad, lo visual y lo intuitivo que es manejar la transformación de datos.
- Si un usuario busca la Transformación de datos, en este caso por la característica de la herramienta si desea transformar datos se puede dirigir a Pentaho Data Integration o a Talend Data Integration, en este caso no será tan fácil, ni tan intuitivo, ni tan visual, pero se puede ganar en la automatización del proceso de transformación de datos que se puede dejar como un proceso estándar en caso de que una tarea se repita periódicamente, adicionando que para el caso de TDI se podrá documentar cada paso en el flujo de la transformación ya sea de forma técnica o funcional.

- En conclusión, el usuario puede usar siempre la herramienta que desee y la que más le guste o este enseñado a usar, pero desde lo investigado no hay una recomendación única para el proceso de Transformación de datos, todo depende de lo que el usuario este buscando solucionar, en este caso solo se está sugiriendo que puede utilizar en un caso dado.

Cargar datos (Data Load)

- Si lo que un usuario busca es Cargar datos, por sus característica y forma fácil que permite comunicarse no solo con la nube, sino con las bases de datos y archivos planos se puede dirigir a Talend Data Integration, que puede brindar una capacidad más completa a la hora de cargar datos a distintas fuentes, seguido por Pentaho Data Integration que también cuenta con una gran capacidad de cargar datos a distintas fuentes.
- Si lo que un usuario busca es Cargar datos, por su característica OpenRefine no es una herramienta muy adecuada para la carga de datos ya que no posee componentes que le permitan cargar datos a distintas fuentes que ahora en día son esenciales como bases de datos y la nube, entre otras, por ese motivo no es una buena opción para tenerla en cuenta para este proceso, se deberá esperar nuevas versiones para saber si se mejora en este componente.
- En conclusión, el usuario puede usar siempre la herramienta que desee y la que más le guste o este enseñado a usar, pero desde lo investigado no hay una recomendación única para el proceso de Carga de datos, todo depende de lo que el usuario este buscando solucionar, en este caso solo se está sugiriendo que puede utilizar en un caso dado.

10 Trabajos futuros

El presente trabajo final logró un acercamiento de las herramientas de ETL, PDI, TDI y OR, con las cuales se puede facilitar el proceso de extracción, transformación y carga de datos de múltiples fuentes. Aunque solo se comparó las herramientas con varios formatos, quedando mucho por cubrir. Por consiguiente, se listan posibles trabajos futuros:

- Comparar las herramientas con conjuntos de datos, que posean formatos de datos semi-estructurados como (XML, JSON, entre otros).
- Comparar las herramientas con conjuntos de datos, que no posean una estructura, es decir con formatos no estructurados, que es donde se ve la mayoría de información como son los datos de redes sociales, textos, emails, datos de la web.
- Comparar e identificar cuál de las herramientas se adapta mejor al mundo de Big Data, que es donde se combinan los datos estructurados, semi-estructurados y no estructurados, con el objetivo de poder procesar y transformar los datos en información útil para la toma de decisiones.

Glosario de términos

- **Lesser General Public License (LGPL)**: Es la Licencia Pública General Reducida de GNU.
- **Java Database Connectivity (JDBC)**: Es una API que permite la ejecución de operaciones sobre bases de datos desde el lenguaje de programación **Java**.
- **Big Data**: Grandes volúmenes de datos
- **Knowledge Discovery on Database**: Descubrimiento de conocimiento en bases de datos
- **CRM**: Customer Relationship Management
- **ERP**: Enterprise Resource Planning
- **SCM**: Supply Chain Management
- **ETL**: Extraction, Transformation and Load
- **Startup**: Empresa emergente que apenas está empezando en el mercado
- **Data Warehouse (DW)**: Bodega de Datos
- **Business Intelligence (BI)**: Inteligencia de Negocios
- **Data Mining (DM)**: Minería de Datos
- **JSON**: JavaScript Object Notation
- **XML**: eXtensible Markup Language
- **RDF**: Resource Description Framework
- **TSV**: Tab Separated Values

11 Referencias bibliográficas

[1] I. Figini and R. G. Illescas, “Tesis de Grado Integrador de fuentes de datos para la gestión por indicadores,” 2018.

[2] C. Walsh, “Data and Analytics: Open Source Data Integration Tool Comparison,” 2016.

[3] D. Server, D. Px, T. O. Studio, and P. D. Integrator, “ETL Benchmarks,” *Informatica*, pp. 1–140, 2008.

[4] T. Input, T. Output, X. M. L. Output, and M. Output, “Benchmark: Talend Open Studio vs Pentaho Data Integrator (aka Kettle) V0.23,” *Chart*, pp. 2–18, 2007.

[5] “GRADO EN SISTEMAS DE INFORMACIÓN Autor: Jesús Melchor Ballesteros Tutor/es: Iván González Diego 2016/2017,” 2017.

[6] “2018 ETL Tools Comparison - DZone Big Data.” [Online]. Available: https://dzone.com/articles/2018-etl-tools-comparison-1. [Accessed: 10-Aug-2018].

[7] “Comparativa Kettle (Pentaho Data Integration) y Talend ~ Business Intelligence y Big Data: ¡Aprende Gratis sobre Analytics!” [Online]. Available: http://www.todobi.com/2016/04/comparativa-kettle-pentaho-data.html. [Accessed: 10-Aug-2018].

[8] “Open Source ETL Tools Comparison | Alooma.” [Online]. Available: https://www.alooma.com/blog/open-source-etl-tools-comparison. [Accessed: 10-Aug-2018].

[9] “Comparativa herramientas ETL.” [Online]. Available: https://es.slideshare.net/JorgeCarlos3/comparativa-herramientas-etl. [Accessed: 10-Aug-2018].

[10] “Integración de datos: Concepto e importancia en la empresa actual.” [Online]. Available: https://www.powerdata.es/integracion-de-datos. [Accessed: 02-Oct-2018].

[11] W. Paper, “Data Integration Déjà Vu : Title Big Data Reinvigorates DI ii.”

[12] W. Paper, “Una Nueva Clase de BI : Analítica de Autoservicio que le Encantará a su Negocio y a sus Usuarios Contenido.”

[13] D. Pianko, “Analytics to Fight Tax Fraud,” no. March, 2016.

[14] A. Bustamante Martínez, E. A. Galvis Lista, and L. C. Gómez Flórez, “Técnicas de modelado de procesos de ETL: una revisión de alternativas y su aplicación en un proyecto de desarrollo de una solución de BI. (Spanish),” *ETL Process. Model. Tech. an Altern. Rev. its Appl. a BI Solut. Dev. Proj.*, vol. 18, no. 1, pp. 185–191, 2013.

[15] “Extract, transform and load - Wikipedia, la enciclopedia libre.” [Online]. Available: https://es.wikipedia.org/wiki/Extract,_transform_and_load. [Accessed: 11-May-2018].

[16] J. P. A. Runtuwene, I. R. H. T. Tangkawarow, C. T. M. Manoppo, and R. J. Salaki, “A Comparative Analysis of Extract, Transformation and Loading (ETL) Process,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 306, no. 1, 2018.

[17] Inmon, *Building the Data Warehouse*, Third Edit. 2002.

[18] M. Valdiviezo, I. Herrera, and G. Jáuregui, “Análisis y Diseño de una herramienta de

desarrollo de soluciones para inteligencia de negocios - Análisis dimensional," *Test*, pp. 1–125, 2007.

[19] A. Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, S. J., Hidalgo-Troya, AlvaradoPérez, and J. C., "The Process of Knowledge Discovery on Databases," no. 2016, pp. 63–86.

[20] T. Morgado-García, D. Antonio, P.-D.-L.-L. Ii, and A. Rosete, "Knowledge discovery in databases historical of a trading company."

[21] J. H. Cáceres, "Descubrimiento de conocimiento en la base de datos académica de una institución de educación superior usando redes neuronales Knowledge discovery in the academic database of a higher education institution using neural networks," 2011.

[22] A. C. Vera García, "Análisis de herramientas BI en el mercado actual," pp. 1–89, 2015.

[23] "Herramientas ETL." [Online]. Available: http://carlosproal.com/dw/dw05.html. [Accessed: 08-Aug-2018].

[24] A. Maria, I. Florea, V. Diaconita, and R. Bologa, "Data integration approaches using ETL," *Database Syst. J.*, vol. VI, no. 3, pp. 19–27, 2015.

[25] M. N. Mali and M. S. Bojewar, "A Survey of ETL Tools," *Int. J. Comput. Tech.*, vol. 2, no. 5, pp. 20–27, 2015.

[26] R. Mukherjee, "A Comparative Review Of Data Warehousing ETL Tools With New Trends And Industry Insight," *2017 IEEE 7th Int. Adv. Comput. Conf.*, pp. 944–949, 2017.

[27] "Cuadrante Mágico de Gartner 2018 de Herramientas de Integración de Datos | Denodo." [Online]. Available: https://www.denodo.com/es/pagina/cuadrante-magico-de-gartner-2018-de-herramientas-de-integracion-de-datos. [Accessed: 04-Aug-2018].

[28] "Comparativa ETL´s OpenSource vs ETL´s Propietarias | respinosamilla blog." [Online]. Available: http://www.dataprix.com/blogs/respinosamilla/comparativa-etl-s-opensource-vs-etl-s-propietarias. [Accessed: 08-Aug-2018].

[29] "Comparando soluciones ETL Open Source y comerciales by Josep Curto Díaz - BeyeNETWORK Edición Español." [Online]. Available: http://www.beyenetwork.es/view/9013. [Accessed: 08-Aug-2018].

[30] "Herramientas ETL y su relevancia en la cadena de valor del dato - Deusto Data." [Online]. Available: https://blogs.deusto.es/bigdata/herramientas-etl-y-su-relevancia-en-la-cadena-de-valor-del-dato/. [Accessed: 10-Aug-2018].

[31] "Proceso y herramientas ETL." [Online]. Available: https://etl-tools.info/es/bi/proceso_etl.htm. [Accessed: 10-Aug-2018].

[32] "11. Herramientas ETL. ¿Que son, para que valen?. Productos mas conocidos. ETL´s Open Source. « El Rincon del BI." [Online]. Available: https://churriwifi.wordpress.com/2009/12/29/11-herramientas-etl-¿que-son-para-que-valen-productos-mas-conocidos-etl´s-open-source/. [Accessed: 10-Aug-2018].

[33] M. Novak and K. Rabuzin, "Prototype of a Web ETL Tool," *Int. J. Adv. Comput. Sci. Appl.*, vol. 5, no. 6, pp. 97–103, 2014.

[34] I. J. Of, "Research in Computer Applications and Robotics Issn 2320-7345 Etl Tools in Data Mining a Review," vol. 2, no. 1, pp. 62–69, 2014.

[35] “Jaspersoft® ETL | Jaspersoft Community.” [Online]. Available: https://community.jaspersoft.com/project/jaspersoft-etl/releases. [Accessed: 15-Aug-2018].

[36] Pentaho, “Data Integration - Kettle | Hitachi Vantara Community.” [Online]. Available: https://community.hitachivantara.com/docs/DOC-1009855-data-integration-kettle. [Accessed: 05-May-2018].

[37] Pentaho, “Integración de datos Plataforma | Pentaho,” 2016. [Online]. Available: http://www.pentaho.com/product/data-integration. [Accessed: 12-Apr-2016].

[38] Pentaho, “Manual del Usuario de Spoon - Pentaho Data Integration (Spanish) - Pentaho Wiki,” 2016. [Online]. Available: http://wiki.pentaho.com/display/EAIes/Manual+del+Usuario+de+Spoon. [Accessed: 12-Apr-2016].

[39] Pentaho, “Integración de Datos | Comunidad Pentaho,” 2016. [Online]. Available: http://community.pentaho.com/projects/data-integration/. [Accessed: 12-Apr-2016].

[40] Talend, “Welcome to Talend Help Center.” [Online]. Available: https://help.talend.com/. [Accessed: 05-May-2018].

[41] “Talend Real-Time Big Data Integration Tools for MDM and ETL.” [Online]. Available: https://www.talend.com/. [Accessed: 11-May-2018].

[42] Talend, “Talend Data Integration Software: integración de datos empresariales,” 2016. [Online]. Available: https://www.talend.com/products/data-integration. [Accessed: 10-Nov-2016].

[43] Talend, “Software de integración de datos en tiempo real Talend Open Source,” 2016. [Online]. Available: https://www.talend.com/. [Accessed: 10-Nov-2016].

[44] “Google Official Blog: Deeper understanding with Metaweb.”

[45] “From Freebase Gridworks to Google Refine and now OpenRefine.”

[46] “OpenRefine.” [Online]. Available: http://openrefine.org/. [Accessed: 10-Apr-2018].

[47] “Herramientas ETL. ¿Que son, para que valen?. Productos mas conocidos. ETL´s Open Source. | respinosamilla blog.” [Online]. Available: http://www.dataprix.com/blogs/respinosamilla/herramientas-etl-que-son-para-que-valen-productos-mas-conocidos-etl-s-open-sour. [Accessed: 10-Aug-2018].

[48] A. Mesa, C. Lochmuller, and M. S. Tabares, “Comparativo entre herramientas BPMN,” *Rev. Soluciones Postgrado*, vol. 6, no. 12, pp. 95–108, 2014.

[49] Universidad, “Método de evaluación y selección de herramientas de simulación de redes,” *Rev. Sist. y Telemática*, vol. 9, no. 16, pp. 55–71, 2011.

[50] Leonel Alfonso Villamizar Gutiérrez, “Cómo abordar un proyecto de Business Intelligence en una empresa u organización,” p. 107, 2010.

[51] “Inicio.” [Online]. Available: https://www.dane.gov.co/. [Accessed: 25-Aug-2018].