AYI ACADEMY: TALEND

Material práctico



Módulo 1: Working with Files

Trabajando con archivos: tratamiento de cada tipo de datos

Este módulo le enseña cómo manejar diferentes tipos de archivos estructurados y cómo realizar operaciones básicas en los archivos (leer, copiar, eliminar y enumerar). Se centra en el uso de componentes de Talend Studio dedicados a la lectura de datos de archivos estructurados (formato CSV) y archivos jerárquicos (formatos JSON y XML).

Abriendo el proyecto

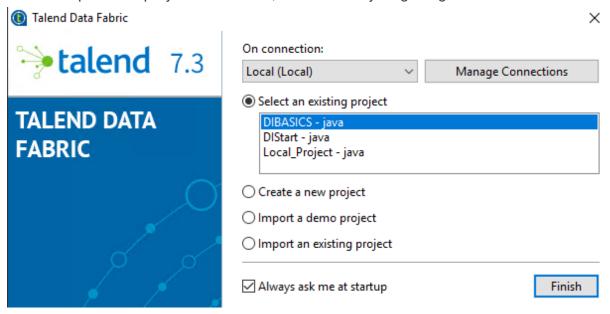
1) Haga clic en el icono de Studio para iniciar Talend Studio.



2) Abra el proyecto DIBASICS.

Confirme que el proyecto DIBASICS está disponible en la lista de proyectos existentes.

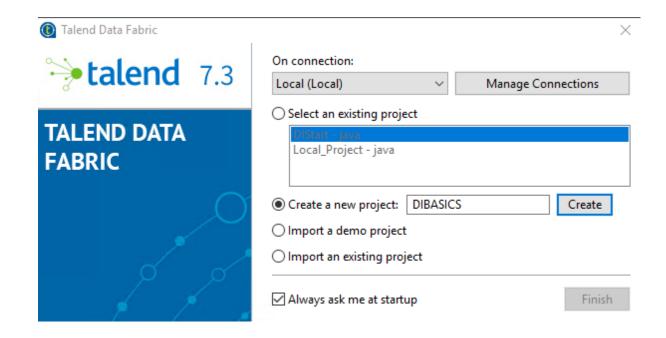
Si aparece el proyecto DIBASICS, selecciónelo y luego haga clic en Finalizar.



Si falta el proyecto, siga estos pasos para crearlo:

Para **On connection**, seleccione **Local (Local)**, lo que significa que cuando se crea, el proyecto se almacena localmente, en su máquina.

Haga clic en el botón de **Create a new project** e ingrese DIBASICS.

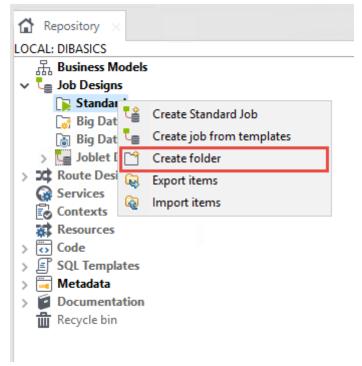


Configura tu Proyecto

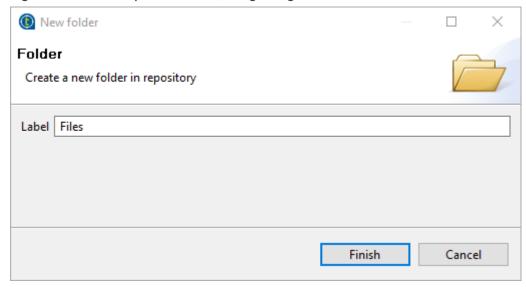
1) En la barra de herramientas de acceso rápido, asegúrese de que la perspectiva de **Integration** esté seleccionada.



2) Cree una nueva carpeta para sus jobs. En el Repositorio, expanda **Job Designs**, haga clic con el botón derecho en **Standard** y seleccione **Create folder**.



Ingrese "archivos" para la Label, luego haga clic en Finalizar.



La nueva carpeta aparece en **Repository** > **Job Designs** > **Standard**.

Este ejercicio le enseña cómo leer e interpretar datos provenientes de un archivo delimitado.

Al final, el Job se ve así. Lee datos de un archivo CSV y los imprime en la consola.



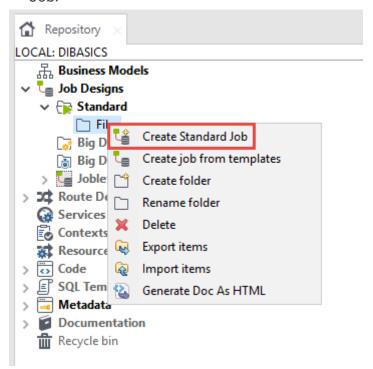
Lectura de datos de un archivo delimitado

En este ejercicio, utilizará el archivo Customers.csv ubicado en la carpeta C:\StudentFiles\DIBasics\Files como ejemplo. Examine el contenido de este archivo, como se muestra a continuación, y observe que cada línea contiene información básica del cliente, separada por comas. La primera línea es un encabezado que contiene los nombres de las columnas de datos.

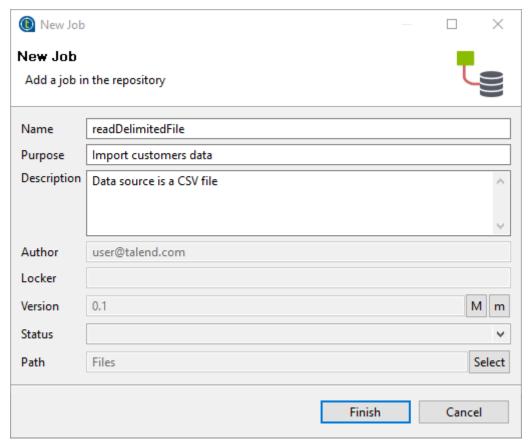
```
Customers.csv 🔣
     First, Last, Number, Street, City, State
     Bill, Coolidge, 85013, Via Real, Austin, IL
  2
     Thomas, Coolidge, 63489, Lindbergh Blvd, Springfield, ca
  3
     Harry, Ford, 97249, Monroe Street, Salt Lake City, ca
     Warren, McKinley, 82589, Westside Freeway, Concord, ak
     Andrew, Taylor, 29886, Padre Boulevard, Madison, CA
  7
     Ulysses, Coolidge, 98646, Bayshore Freeway, Columbus, MN
  8
     Theodore, Clinton, 12292, San Marcos, Bismarck, NY
  9
     Benjamin, Jefferson, 82077, Carpinteria North, Sacramento, ca
     William, Van Buren, 21712, Tully Road East, Albany, IL
 10
     Calvin, Washington, 50742, Richmond Hill, Charleston, ca
 11
 12
     Jimmy, Polk, 76143, Richmond Hill, Salt Lake City, AK
     Calvin, Adams, 52386, Lake Tahoe Blvd., Montgomery, NY
 13
     Ulysses, Monroe, 70511, Jones Road, Trenton, IL
 14
     Zachary Tyler, 45040, Santa Rosa North, Carson City, AK
     Ulysses, Johnson, 19989, Via Real, Juneau, AL
```

Utilice un componente tFileInputDelimited para leer un archivo delimitado fila por fila, dividir datos en campos y enviar los campos, como se define en el esquema, al siguiente componente.

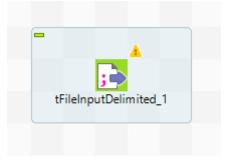
- 1) Cree un nuevo trabajo en la carpeta Archivos y asígnele el nombre readDelimitedFiles.
 - a) En el **Repository**, expanda **Job Designs** > **Standard** y haga clic con el botón derecho en **Files**.
 - b) Seleccione Create Standard Job. Se abre el cuadro de diálogo New Job.



- c) En el cuadro de texto **Name**, ingrese readDelimitedFile.
- d) Rellene los cuadros de texto **Purpose** y **Description** como se muestra.
- e) Haga clic en Finalizar. El Trabajo se abre en el Designer.

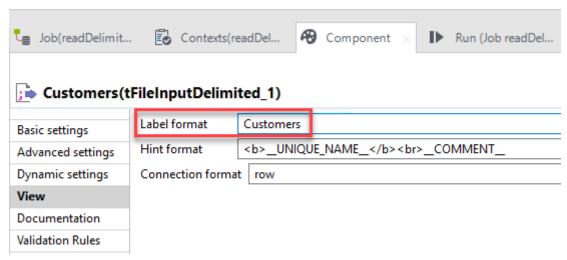


- 2) Agregue un componente que lea desde un archivo delimitado.
 - a) Haga clic en cualquier lugar en el Designer.
 - b) Comience a escribir el nombre del componente,
 tFileInputDelimited. A medida que escribe, aparece una lista de componentes coincidentes.
 - c) Haga doble clic en tFileInputDelimited en la lista. El componente aparece en el Diseñador.



- 3) Cambie el nombre del componente a Clientes.
 - a) Haga clic en la pestaña View en la Component view.
 - b) Reemplace el texto en el cuadro de texto **Label format** con Clientes.

Esto actualiza automáticamente el nombre del componente en Designer.

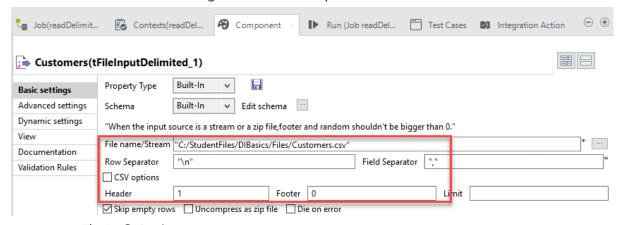


4) Configure la configuración básica de componentes.

Para configurar el componente Clientes, especifique las propiedades del archivo, como el nombre del archivo para leer, los separadores de filas y campos. Los datos se leen del archivo fila por fila, utilizando el separador de fila para identificar el final de una línea. Cada fila se separa a su vez en campos utilizando el separador de campos y se interpreta de acuerdo con el esquema.

- a) En la vista de componentes, abra la pestaña Basic settings.
- b) Para seleccionar el **File name**, haga clic en [...] y seleccione la ruta del archivo C:/StudentFiles/DIBasics/Files/Customers.csv.
- c) Deje el **Row Separator** predeterminado como "\n", que corresponde al final de la línea.
- d) Para especificar que los campos de una fila estén separados por comas, en **Field Separator**, ingrese ","
- e) En el cuadro de texto **Header**, ingrese 1. El componente ignora la primera línea del archivo, que no contiene datos reales, solo los nombres de las columnas de datos.

La configuración del componente se ve así:

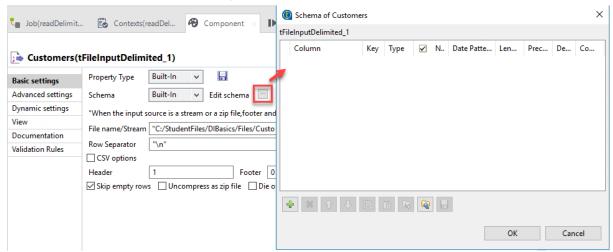


5) Definir el esquema

Por cada línea leída en el archivo, el componente asigna cada campo de datos a una columna de su esquema definido. Para cada campo de datos

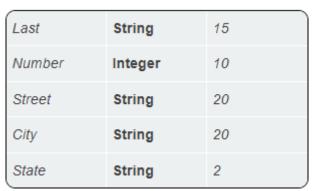
extraído del archivo, el esquema necesita una columna correspondiente con un nombre y atributos, como tipo y longitud.

a) Para definir manualmente el esquema, a la derecha de **Edit schema**, haga clic en [...]

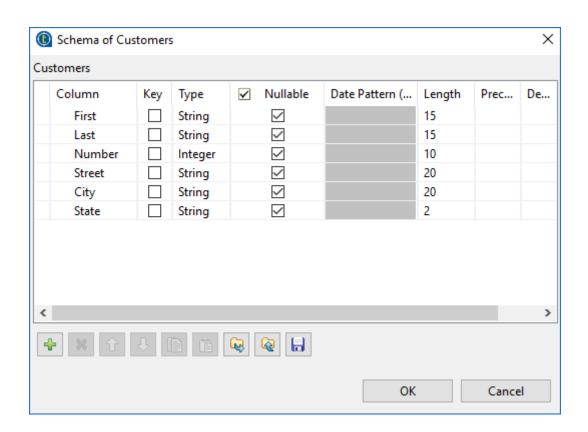


- b) Haga clic en [+] para agregar una nueva columna al esquema.
 Asigne el nombre Primero a la columna. Para Type, seleccione
 String y para Length, ingrese 15.

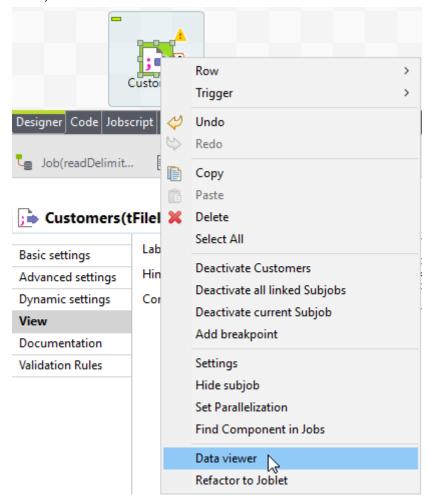
 Esto significa que la primera columna de cada fila del archivo contiene un valor de String con un máximo de 15 caracteres que representan el nombre del cliente.
- Repita estos pasos para agregar el resto de las columnas a su esquema. Seleccione los tipos y longitudes de la siguiente manera:



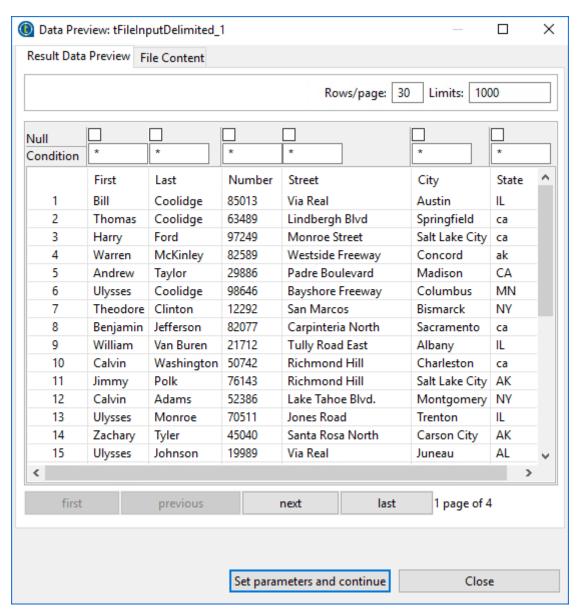
Confirme que la definición del esquema es correcta como se muestra a continuación y haga clic en Aceptar.



- 6) Utilice el visor de datos del componente Clientes para obtener una vista previa de los datos extraídos del archivo.
 - a) Haga clic con el botón derecho en el componente Clientes.
 - b) Seleccionar Data viewer



Una ventana muestra una vista previa de los datos recuperados del archivo Customers.csv. Esta es una buena manera de confirmar que la configuración del lector de archivos se configuró correctamente. La pestaña Contenido del archivo muestra el contenido sin procesar del archivo.

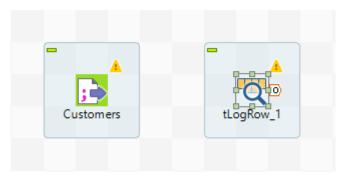


Click Close.

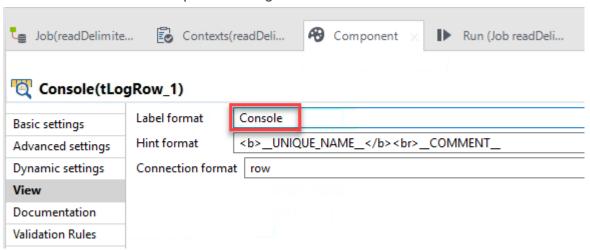
Impresión del contenido del archivo en la consola

Para mostrar el contenido leído del archivo en la consola, conecte el componente de entrada del archivo a un componente de salida tLogRow.

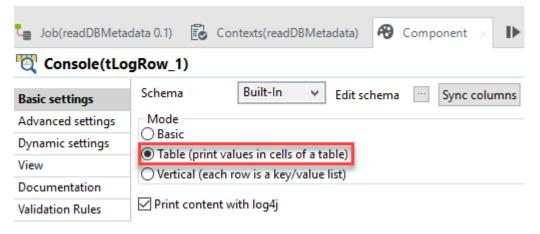
1) Agregue un componente tLogRow a la derecha del componente Clientes.



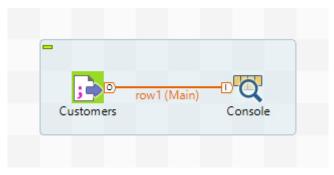
2) Cambie el nombre del componente tLogRow a Consola.



- 3) Configure el componente tLogRow para imprimir los resultados en una tabla.
 - a) En la vista de componentes, abra la pestaña Basic settings.
 - b) Para **Mode**, seleccione **Table**, que le permite imprimir los resultados en una tabla.

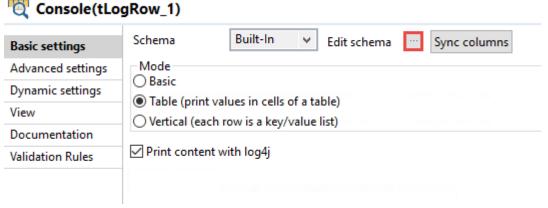


- 4) Conecte los componentes Clientes y Consola mediante una fila Principal.
 - a) Haga clic con el botón derecho en el componente de entrada, Clientes.
 - b) Select **Row** > **Main**. A connection link is visible between the input component and mouse pointer.
 - c) Haga clic en el componente **Console** para crear el enlace.

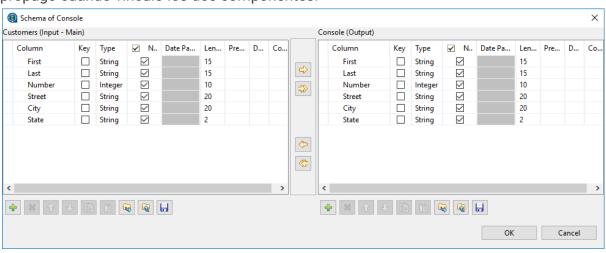


tip: Las conexiones de fila iteran en filas secuencialmente y transfieren datos entre componentes. La fila principal es el tipo de conexión de fila más común y se usa de forma predeterminada.

 Revisa el esquema.
 Seleccione el componente tLogRow y haga clic en el botón Edit schema ([...]) en la vista de componentes.



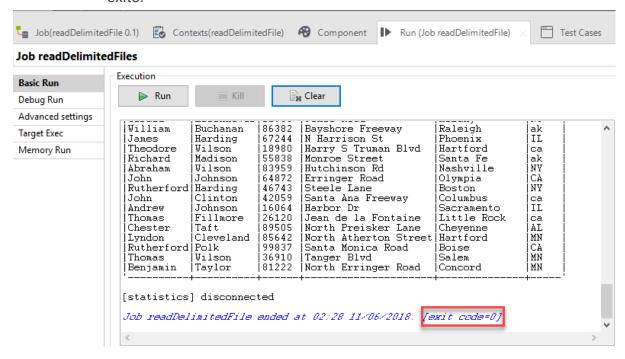
Observe que el esquema que especificó en el componente tFileInputDelimited se propagó cuando vinculó los dos componentes.



Click OK.

- 6) Ejecute el trabajo.
 - a) Abra la vista **Run** debajo del **Designer**.
 - b) Haga clic en **Run** . Se ejecuta el Job y se muestra en la consola la tabla que contiene la lista de clientes extraída del archivo. Confirme que el

trabajo termina con **[exit code = 0]**, lo que significa que se ejecutó con éxito.



Este ejercicio le enseña cómo leer e interpretar datos provenientes de archivos jerárquicos.

Al final, los dos trabajos se ven así. El primer trabajo lee datos de un archivo XML y los imprime en la consola. De manera similar, el segundo trabajo lee datos de un archivo JSON y los imprime en la consola.



En este ejercicio, utilizará los archivos ProductSales.xml y ShopSales.json ubicados en la carpeta C:\StudentFiles\DIBasics\Files.

Lectura de datos de archivos XML jerárquicos

El primer trabajo tiene que leer datos del archivo ProductSales.xml. Mire el contenido del archivo (que se muestra a continuación) para comprender su estructura. La etiqueta raíz, como lo indica su nombre, es el elemento superior del documento XML y contiene todas las etiquetas de fila. Observe que las etiquetas de fila agrupan datos sobre ventas de productos, como el nombre de la tienda, la cantidad, el nombre del cliente y el producto.

```
<root>
    <row>
         <ShopName>Shop1</ShopName>
         <Quantity>790</Quantity>
         <CustomerName>Acme Chemical/CustomerName>
         <Product>Product 34</Product>
    </row>
    <row>
         <ShopName>Shop1</ShopName>
         <Quantity>886</Quantity>
         <CustomerName>Truck-X Repair
         <Product>Product 36</Product>
    </row>
    <row>
         <ShopName>Shop1</ShopName>
         <Quantity>972</Quantity>
         <CustomerName>Lee Stairworks</CustomerName>
         <Product>Product 47</Product>
    </row>
</root>
```

Para leer datos de un archivo XML, utiliza un componente tFileInputXML.

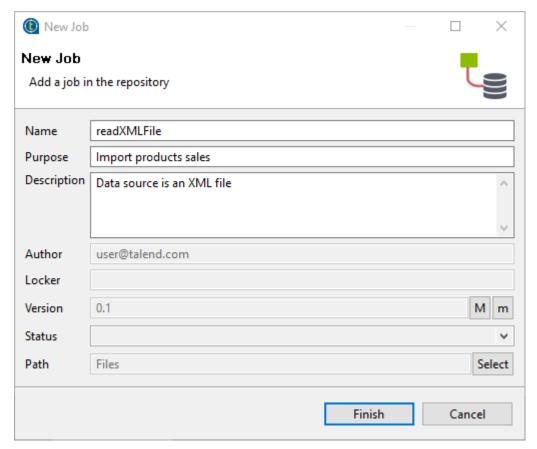
Un componente tFileInputXML extrae los datos almacenados en el archivo XML, los asigna a los campos definidos por el esquema y los transfiere al siguiente componente.

Para seleccionar elementos y atributos en documentos XML, utilice las expresiones XPath estándar. En este ejemplo, para obtener los nodos que contienen valores, el analizador debe seleccionar, comenzando desde "raíz", los hijos de los nodos de "fila". La expresión XPath correspondiente es /root/row.

1) Crear un trabajo.

En el Repositorio, expanda Diseños de trabajo > Estándar, luego haga clic con el botón derecho en Archivos y cree un nuevo trabajo estándar llamado readXMLFile.

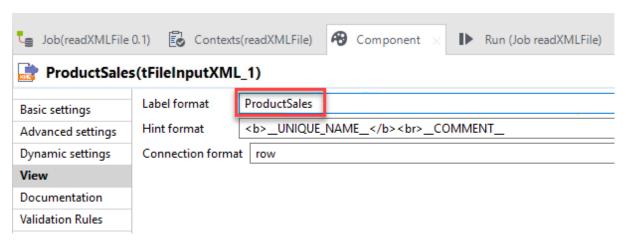
El Trabajo se abre en el Diseñador.



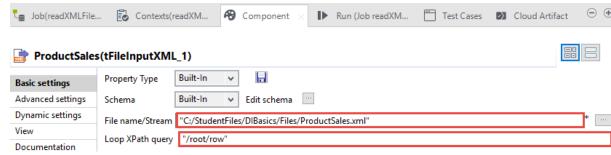
Agregue un componente que lea archivos XML
 Agregue un nuevo componente tFileInputXML al Diseñador.



3) Cambie el nombre del componente tFileInputXML a ProductSales.

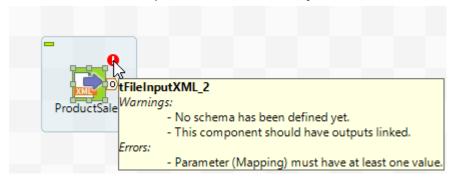


- 4) Configure el componente ProductSales, incluido el nombre del archivo para leer y sus propiedades.
 - a) En la vista de componentes, abra la pestaña Configuración básica.
 - b) Para configurar el nombre del archivo, haga clic en [...] y seleccione la ruta del archivo C:/StudentFiles/DIBasics/Files/ProductSales.xml.
 - c) Ingrese la expresión XPath "/root/row" en el cuadro de consulta Loop XPath para seleccionar los nodos que contienen valores.



5) Inspeccionar advertencias y errores.

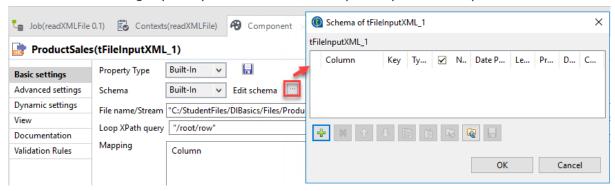
En Designer, observe que el componente ProductSales muestra un error (un icono de signo de exclamación) en la esquina superior derecha. Pase el mouse sobre el icono para mostrar el mensaje.



Un mensaje de advertencia muestra que el esquema no se ha definido. Además, un mensaje de error indica que no se ha realizado la asignación. Esto se debe a que el parámetro Mapeo no está configurado. Este parámetro asocia los campos de datos leídos del archivo con las columnas definidas en el esquema.

6) Cree las columnas del esquema y sus propiedades.

a) A la derecha de Editar esquema, haga clic en [...]. Se abre un cuadro de diálogo que le permite definir un esquema para el componente.

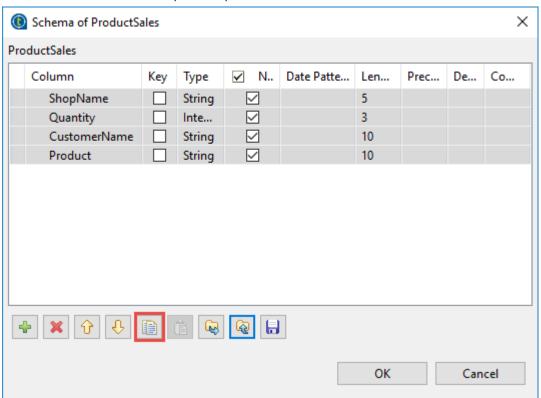


b) Esta vez, carga el esquema desde un archivo local.



En la barra de herramientas, haga clic en la carpeta con el botón de flecha, como se muestra en la captura de pantalla, y seleccione "C:\StudentFiles\DIBasics\Files\Schema.xml".

c) Asegúrese de que todas las columnas estén seleccionadas y haga clic en el botón Copiar esquema.

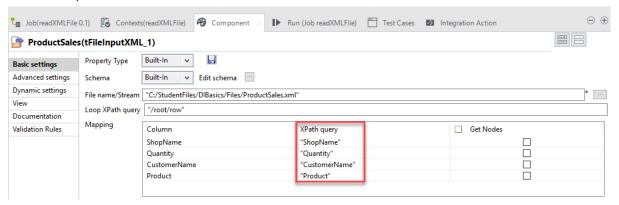


Lo reutiliza más tarde mientras configura el segundo trabajo.

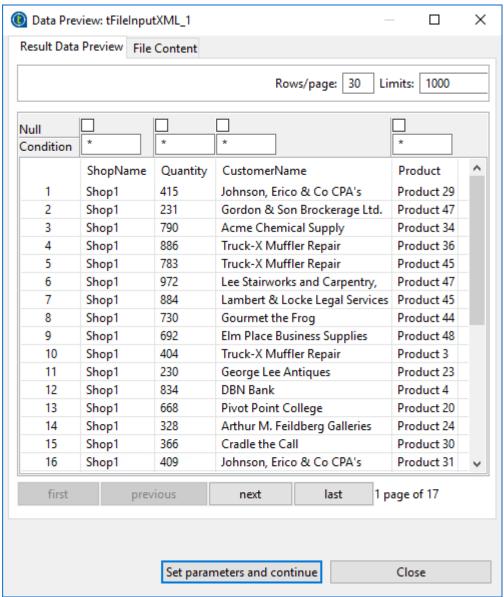
d) Para cerrar la ventana del esquema, haga clic en Aceptar.

En la vista de componentes, en la tabla de asignación, observe que la información de la columna se agrega automáticamente después de definir el esquema.

7) Para configurar la asignación entre los nombres de columna y el contenido real para extraer del archivo, ingrese los nombres de nodo correspondientes exactos utilizados por la estructura xml, entre comillas dobles, como se muestra.



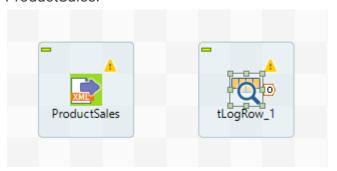
8) Abra el visor de datos en el componente ProductSales haciendo clic con el botón derecho en el componente y luego seleccionando Visor de datos.



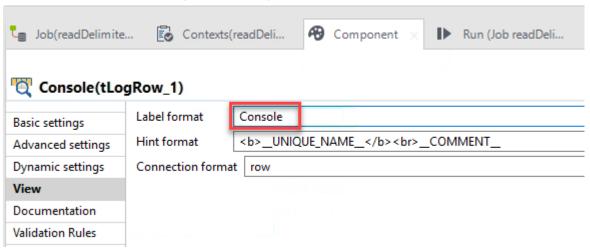
Impresión del contenido del archivo en la consola

Para imprimir el contenido leído del archivo en la consola, conecte el lector de archivos a un componente tLogRow.

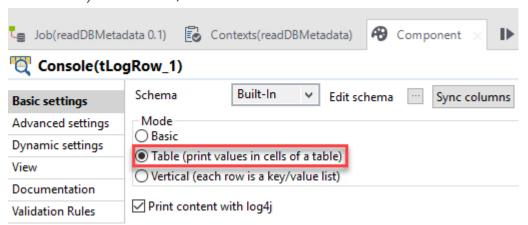
 Agregue un nuevo componente tLogRow a la derecha del componente ProductSales.



2) Cambie el nombre del componente tLogRow a Console.



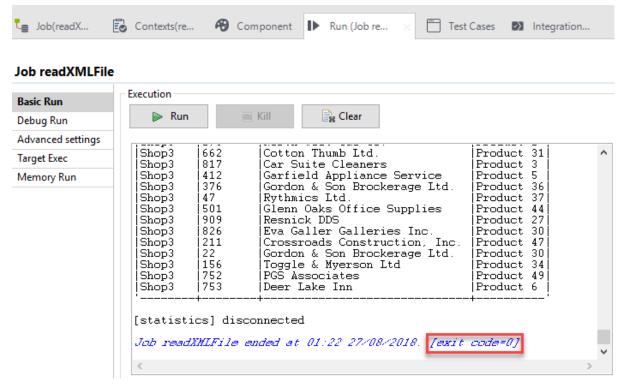
- 3) Configure el componente tLogRow para imprimir los resultados en una tabla.
 - a) En la vista de componentes, seleccione la pestaña Configuración básica.
 - b) Para Modo, seleccione Tabla.



4) Conecte los componentes ProductSales y Console. Haga clic con el botón derecho en el componente de entrada, ProductSales, y use una fila Principal para conectarlo a la Consola.



5) En la vista Ejecutar, haga clic en Ejecutar.



El trabajo se ejecuta y la tabla que contiene los elementos extraídos del archivo se muestra en la consola. Confirme que el trabajo termina con [código de salida = 0], lo que significa que se ejecutó con éxito.

En el siguiente ejercicio, creará un trabajo similar que lee datos de un archivo JSON.

Lectura de datos de archivos JSON jerárquicos

Examine el archivo ShopSales.json, que se muestra a continuación, para comprender su estructura. Tenga en cuenta que contiene el mismo tipo de información que el archivo XML anterior, solo que el formato es diferente. Ubique los campos que le interesan: nombre de la tienda, cantidad, producto y nombre del cliente.

```
"ShopName": "Shop1",
    "Quantity": 415,
    "Product": "Product 29",
    "CustomerName": "Johnson, Erico & Co CPA's"

},

{
    "ShopName": "Shop1",
    "Quantity": 231,
    "Product": "Product 47",
    "CustomerName": "Gordon & Son Brockerage Ltd."
},

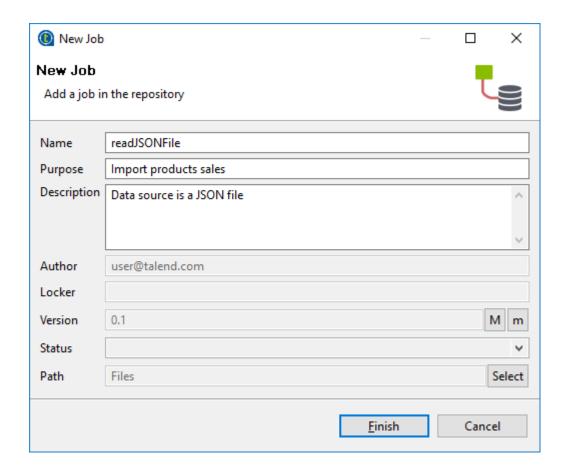
{
    "ShopName": "Shop1",
    "Quantity": 790,
    "Product": "Product 34",
    "CustomerName": "Acme Chemical Supply"
}
```

Para leer datos de un archivo JSON, utiliza un componente tFileInputJSON.

Este componente extrae los datos almacenados en el archivo JSON y los envía a los campos definidos por el esquema, al siguiente componente.

Para este ejercicio, se utiliza el analizador JSONPath. Para configurar el analizador JSONPath, use expresiones JSONPath. JSONPath se comporta de manera muy similar a la forma en que XPath maneja documentos XML. En este ejemplo, para extraer datos sobre ventas, el analizador tiene que seleccionar, desde la parte superior del documento, todos los objetos que son hijos de ShopSales. La expresión JSONPath correspondiente es \$.ShopSales[*]

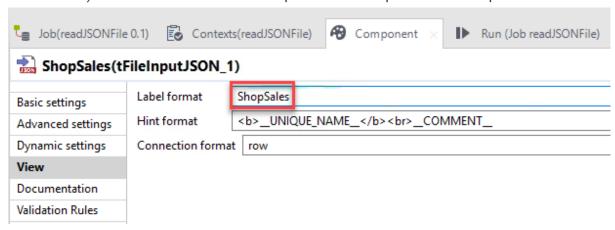
 Crear un trabajo.
 En el Repositorio, expanda Diseños de trabajo > Estándar, luego haga clic con el botón derecho en Archivos y cree un nuevo trabajo llamado readJSONFile.



2) Utilice el mismo método que utilizó anteriormente en este ejercicio para agregar un componente tFileInputJSON en Designer.

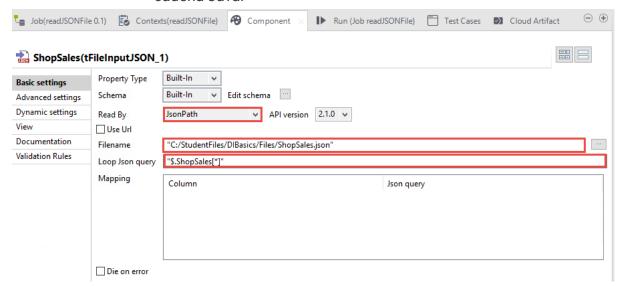


3) Cambie el nombre del componente tFileInputJSON a ShopSales.



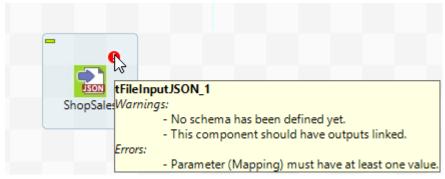
- 4) Configure el componente ShopSales, incluida la ruta del archivo y las opciones del analizador.
 - a) Haga clic en ShopSales para abrir la vista de componentes.
 - b) Para Leer por, seleccione JsonPath y deje la versión de API predeterminada.
 - c) Para seleccionar el nombre del archivo, haga clic en [...] y seleccione la ruta del archivo
 - $\hbox{C:/StudentFiles/DIBasics/Files/ShopSales.json}.$
 - d) Para especificar cómo iterar en el documento para extraer datos, para la consulta Loop Json, ingrese la expresión JSONPath "\$.ShopSales[*]".

Aunque las comillas dobles no forman parte de la expresión, son necesarias porque Studio espera que esta propiedad sea una cadena Java.



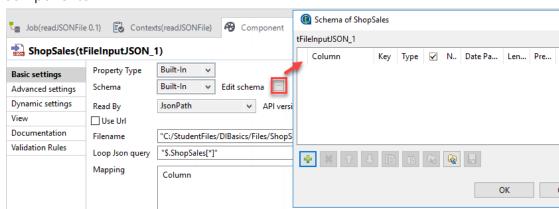
e) En Designer, observe que el componente muestra un error (un icono de signo de exclamación) en la esquina superior derecha.

Pase el puntero del mouse sobre el icono para mostrar el mensaje.



La información sobre herramientas señala que el parámetro Asignación debe tener al menos un valor. Este parámetro indica el mapeo entre los datos leídos del archivo (usando la consulta JSON) y las columnas definidas en el esquema. Para configurar la asignación, debe definir un esquema.

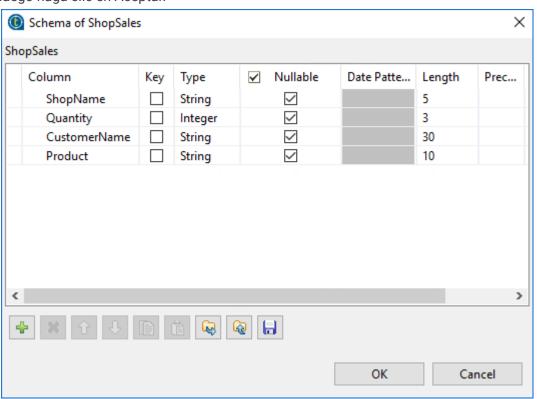
- 5) Cree las columnas del esquema y sus propiedades.
 - a) A la derecha de Editar esquema, haga clic en [...]. Se abre un cuadro de diálogo que le permite definir un esquema para el componente.



b) Para pegar el esquema copiado anteriormente, en la barra de herramientas, haga clic en el botón Pegar.



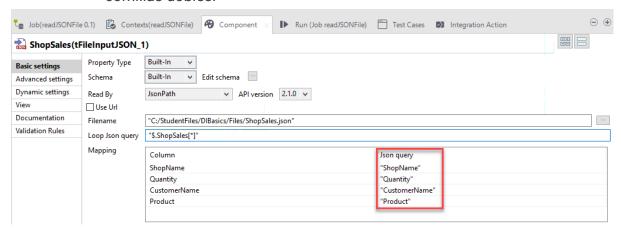
c) Confirme que su esquema esté definido como se muestra a continuación, luego haga clic en Aceptar.



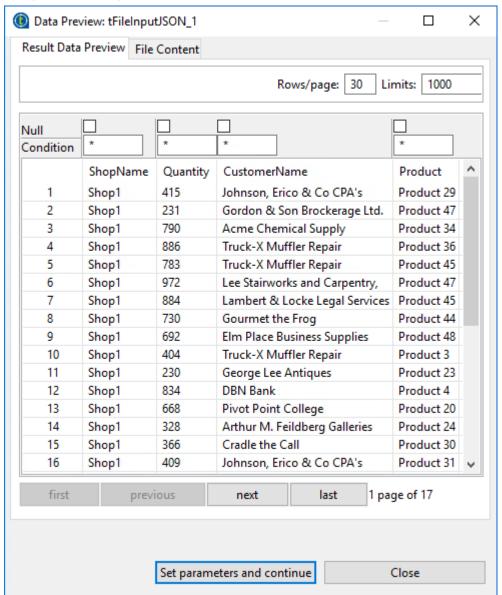
En la sección Asignación, observe que la información de la columna se agrega automáticamente después de definir el esquema.

6) Para configurar la asignación entre los nombres de las columnas y los campos de datos extraídos del archivo, ingrese los nombres de los

campos que se muestran en la siguiente captura de pantalla entre comillas dobles.



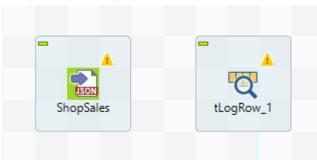
7) Confirme la configuración de asignación abriendo el visor de datos en el componente ShopSales.



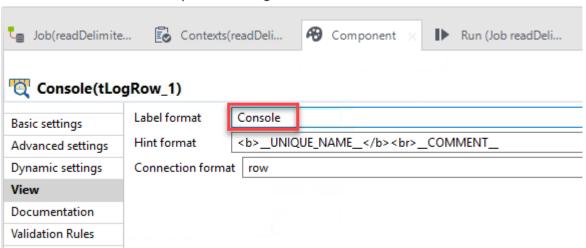
Impresión del contenido del archivo en la consola

Para imprimir el contenido leído del archivo en la consola, conecte el lector de archivos a una consola.

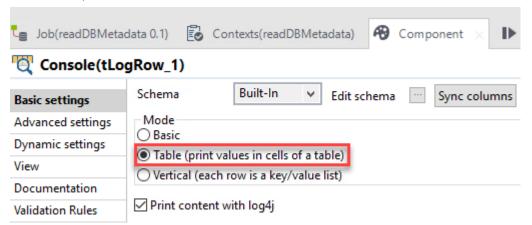
1) En Designer, agregue un nuevo componente tLogRow a la derecha del componente ShopSales.



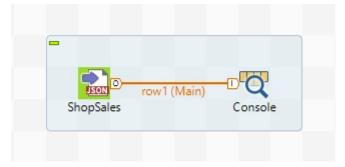
2) Cambie el nombre del componente tLogRow a Console.



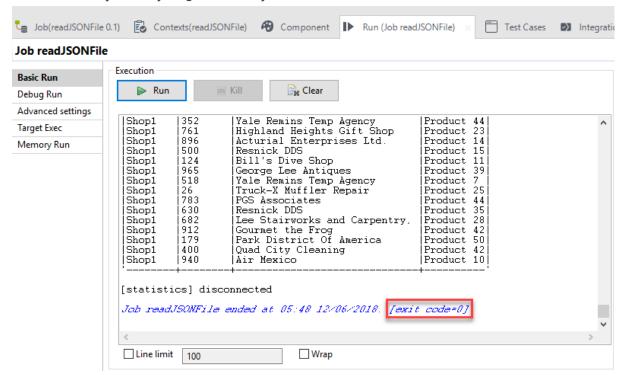
- 3) Configure el componente tLogRow para imprimir los resultados en una tabla.
 - a) Haga clic en Consola para abrir la vista de componentes.
 - b) Para Modo, seleccione Tabla.



4) Conecte ShopSales y el componente Consola mediante una fila Principal.



5) Abra la vista Ejecutar y haga clic en Ejecutar.



El trabajo se ejecuta y la tabla que contiene los elementos extraídos del archivo se muestra en la consola. Confirme que el trabajo termina con [código de salida = 0].