

**UNIVERSIDAD DE COSTA RICA
ESCUELA DE CIENCIAS DE LA
COMPUTACIÓN E INFORMÁTICA**

Introducción a ETL

Elaborado por
Dra. Elzbieta Malinowski G.

Modificado por
Dr. Luis Gustavo Esquivel Quirós

Versión 01-2025



Prácticas Procesos ETL by [Dra. Elzbieta Malinowski Gajda](#)
is licensed under a [Creative Commons Attribution-NonCommercial 3.0 Costa Rica License](#).
Permissions beyond the scope of this license may be available at <http://creativecommons.org>.

OBJETIVO GENERAL DE LA PRÁCTICA

Familiarizarse con la estructura y componentes básicos de la herramienta que permite automatizar los procesos de extracción, transformación y carga (ETL, *extraction-transformation-loading*).

OBJETIVOS ESPECÍFICOS DE LAS PRÁCTICAS

El estudiante aprenderá:

- Extraer los datos de diferentes orígenes (texto, base de datos).
- Transformar los datos de acuerdo con los requerimientos establecidos que permiten su verificación, limpieza, agregación, separación en grupos de datos correctos e incorrectos, entre otros.
- Enviar los datos transformados a diferente tipo de destino (texto, hoja electrónica, base de datos).
- Asegurarse de adecuado tratamiento de los datos incorrectos que no pueden ser procesados.

DOCUMENTACIÓN

La documentación donde cada estudiante debe presentar los resultados de esta práctica debe incluir su nombre, carné, así como las respuestas y pantallazos que demuestren la realización de las actividades. El nombre del o los archivos a entregar deben iniciar por su número de carné, su primer apellido y la palabra PrácticaDirigidaETL, por ejemplo, C07755Esquivel PrácticaDirigidaETL.pdf.

El reporte debe enviarse en formato pdf y además deben incluir el proyecto de Integration Services que realizaron.

METODOLOGÍA

Se especificarán los pasos de ejecución de cada práctica incluyendo los pantallazos en los casos necesarios. Los estudiantes deben realizar la práctica y presentar los proyectos implementados ejecutándolos (a solicitud del docente).

NOTA

En esta práctica se usará SSIS (*SQL Server Integration Services*). Antes de realizarla deben verificar que estén instalados [Integration Services](#) y [Analysis Services](#) (ETL y OLAP de Microsoft) estos paquetes forman parte de Visual Studio y Microsoft SQL Server. Para asegurarse que están corriendo los servicios de SQL Server requeridos para la herramienta SSIS, les recomiendo utilizar la herramienta SQL Server Configuration Manager para verificar que los servicios se encuentren funcionando.



PRÁCTICA DIRIGIDA¹

DESCRIPCIÓN

Se utiliza la base de datos operacional [AdventureWorks2022](#) (distribuida por Microsoft, la cual de no encontrarse en su sistema gestor de bases de datos se debe descargar y restaurar en su motor de base de datos), hacer transformaciones tipo derivación de datos, agregación y ordenamiento y enviar el resultado a un archivo plano y a la hoja Excel.

PASOS:

Previos a la práctica

Documente o explique cada uno de los términos con los que debe familiarizarse.

Familiarizarse con las siguientes opciones del flujo de control:

- Data Flow Task
- Execute SQL Task
- File System Task

Familiarizarse con las siguientes opciones del flujo de datos:

- OLE DB Source
- Derived Column
- Aggregate
- Sort
- Data Viewer
- Multiple destinations
- Flat file destination
- Excel destination

Familiarizarse con:

- Abrir el puerto de la base de datos en el sistema operativo

Durante la práctica

I. Creación del proyecto y paquete

1. Crear un nuevo proyecto de [Integration Services](#) desde Microsoft Visual Studio, llamado PracticaETL1 que debe pertenecer a una carpeta llamada solución PracticasETL. En caso de requerir instalar Microsoft Visual Studio, debe seleccionar los conjuntos de herramientas “Desarrollo de Azure”, “Almacenamiento y procesamiento de datos” y “Desarrollo de Office y SharePoint”. Así como también instalar [Integration Services](#).
2. Cambiar el nombre del paquete o solución por “[SuNombre]AWExtract”.

II. Elementos de control de flujo

¹ Basada en el libro de Brian Knight *et al.*, “Professional SQL Server 2005 Integration Services”, Wiley Publishing, 2006.

En esta práctica se usará solo un elemento: *Data flow*

1. En la pestaña de flujo de control (*Control Flow*) arrastre la tarea de *Data Flow*.
2. Cambiarle el nombre por “SalesReport”.
3. Con doble-clic abrirla para definir sus componentes.

III. Componentes de Data Flow

Orígenes de datos

1. Arrastrar el origen OLE DB Source para el panel de diseño.
2. Cambiarle el nombre por “Transaction History”.
3. Abrir el editor con doble clic y establecer una nueva conexión (si todavía no existe) con la base de datos AdventureWorks2022 de SQL Server.
 - a. Provider: Native OLE DB/SQL Server Native Client 10.0
 - b. Server name: <MAQUINA###> (servidor de SQL Server en ECCI)
 - c. Use la autenticación de Microsoft
 - d. Database name: AdventureWorks2022
4. Para la conexión creada anteriormente:
 - a. Establecer el acceso a la tabla [Production].[TransactionHistoryArchive]
 - b. Entrar a la opción de *Columns* y seleccionar las columnas ProductID, Quantity y ActualCost como se puede ver en la figura 1.

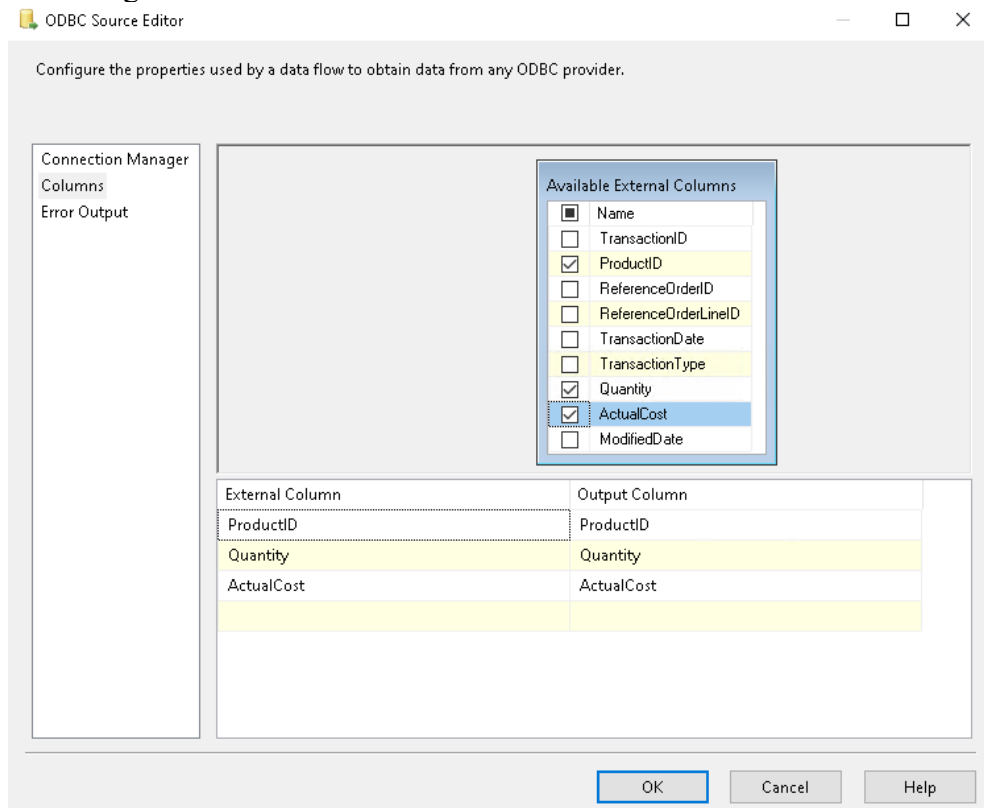
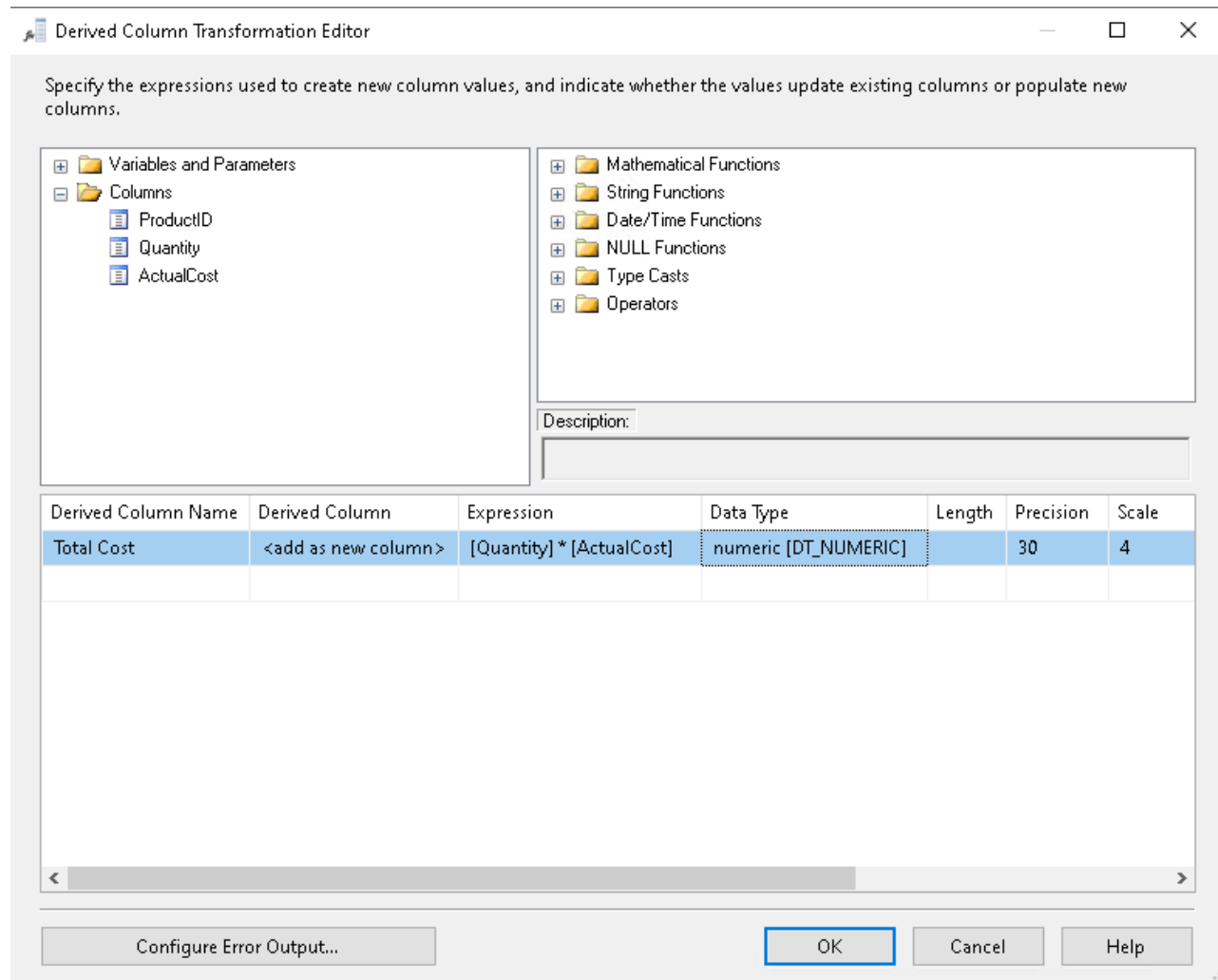


Figure 1 Selección de columnas de datos de origen.

Transformaciones

Columna derivada

1. En el mismo panel de flujo de datos, arrastrar la transformación *Derived Column*, nombrarla “Calculate Total Cost” y conectarla con *sources* usando la flecha azul de origen OLE DB.
2. Con doble clic, abrir el editor de la columna derivada.
3. Escribir el nombre de la columna derivada “TotalCost”, establecerla como una columna nueva y escribir en la columna de *Expression* lo siguiente: $[Quantity] * [ActualCost]$. Debe tener la pantalla como se muestra en la figura 2:

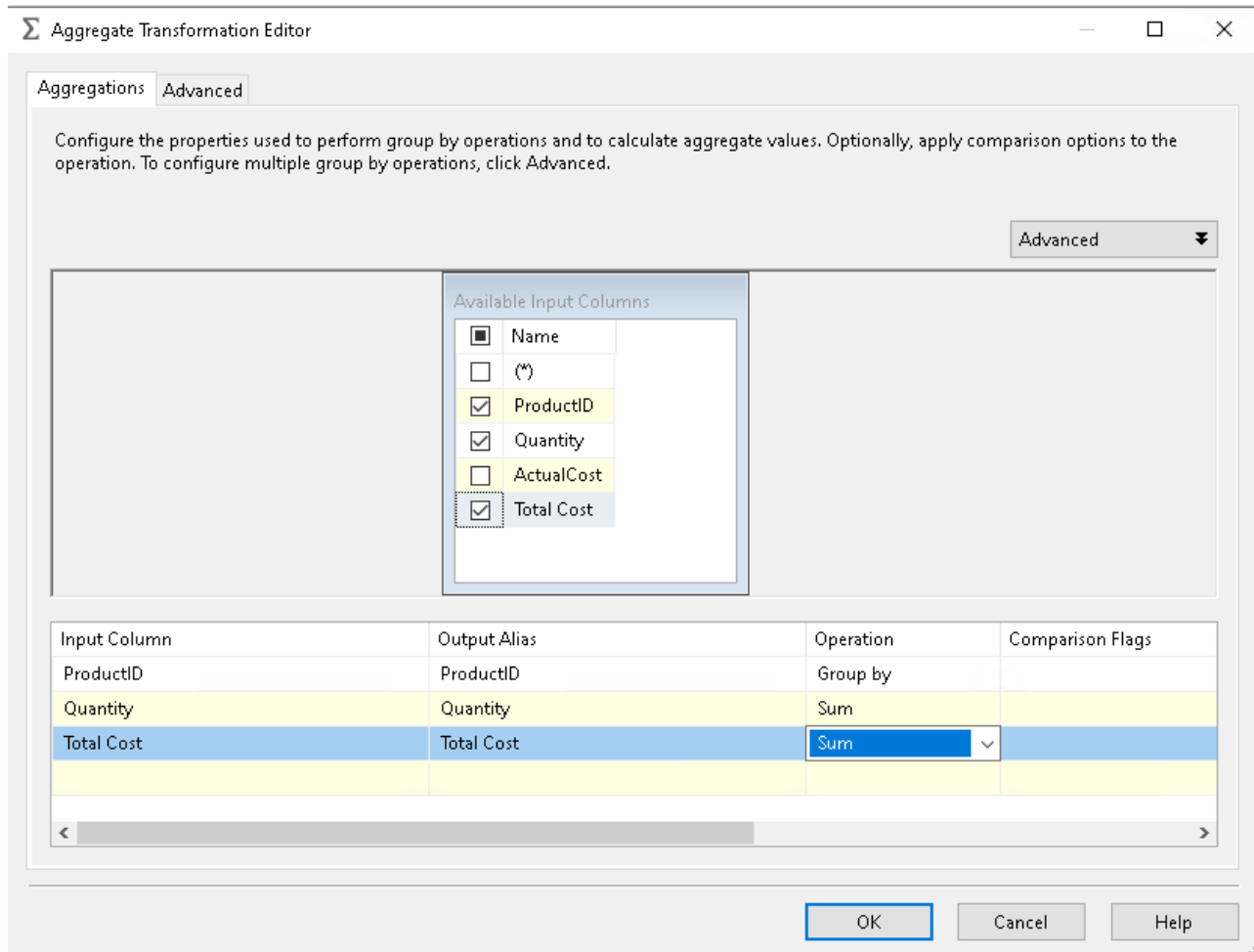


Derived Column Name	Derived Column	Expression	Data Type	Length	Precision	Scale
Total Cost	<add as new column>	[Quantity] * [ActualCost]	numeric [DT_NUMERIC]	30	4	

Figura 2. Pantalla en el editor de la columna derivada

Agregados

1. Arrastrar la opción *Aggregate* y ponerle el nombre “Aggregate Data”. Conectarla con la flecha azul saliente de la transformación *Calculate Total Cost*.
2. Con doble clic, abrir el editor de agregados y seleccionar las columnas (ver la figura 3) de la siguiente manera:
 - a. ProductID con la operación Group by
 - b. Quantity con la operación Sum
 - c. TotalCost con la operación Sum



Aggregate Transformation Editor

Aggregations Advanced

Configure the properties used to perform group by operations and to calculate aggregate values. Optionally, apply comparison options to the operation. To configure multiple group by operations, click Advanced.

Advanced

Available Input Columns

- ☐ Name
- ☐ (*)
- ☒ ProductID
- ☒ Quantity
- ☐ ActualCost
- ☒ Total Cost

Input Column	Output Alias	Operation	Comparison Flags
ProductID	ProductID	Group by	
Quantity	Quantity	Sum	
Total Cost	Total Cost	Sum	

OK Cancel Help

Figura 4. Editor de agregados

Ordenamiento

1. Arrastrar al panel de diseño la transformación de *Sort* y aplicar los pasos parecidos a los anteriores para la conexión con *Aggregate Data*.
2. Establecer lo siguiente en el editor de *Sort*:
 - a. Nombre de la transformación: “Sort by Product Quantity”
 - b. Sort: columna de Quantity
 - c. Sort Type: descendente

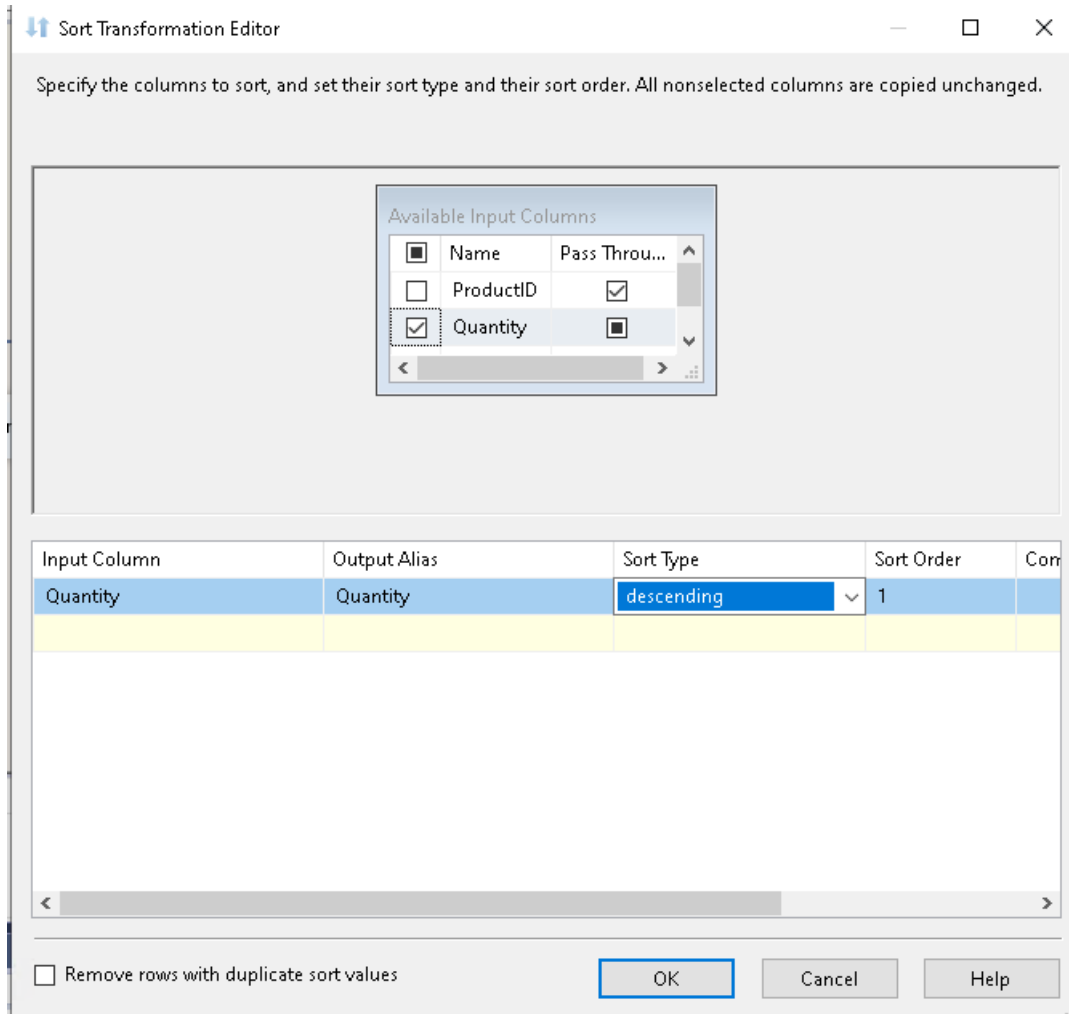


Figura 4. El editor de *Sort*.

Destino

1. Arrastrar el icono de destinación *Flat File Destination* y conectarlo con el *Sort* previamente incluido en *Data Flow*.
2. Con doble clic abrir el editor y establecer una nueva conexión tipo *Delimited*.
3. Especificar los parámetros indicados en la figura 4 (el nombre de archivo de salida es “ProductQtyCost”).

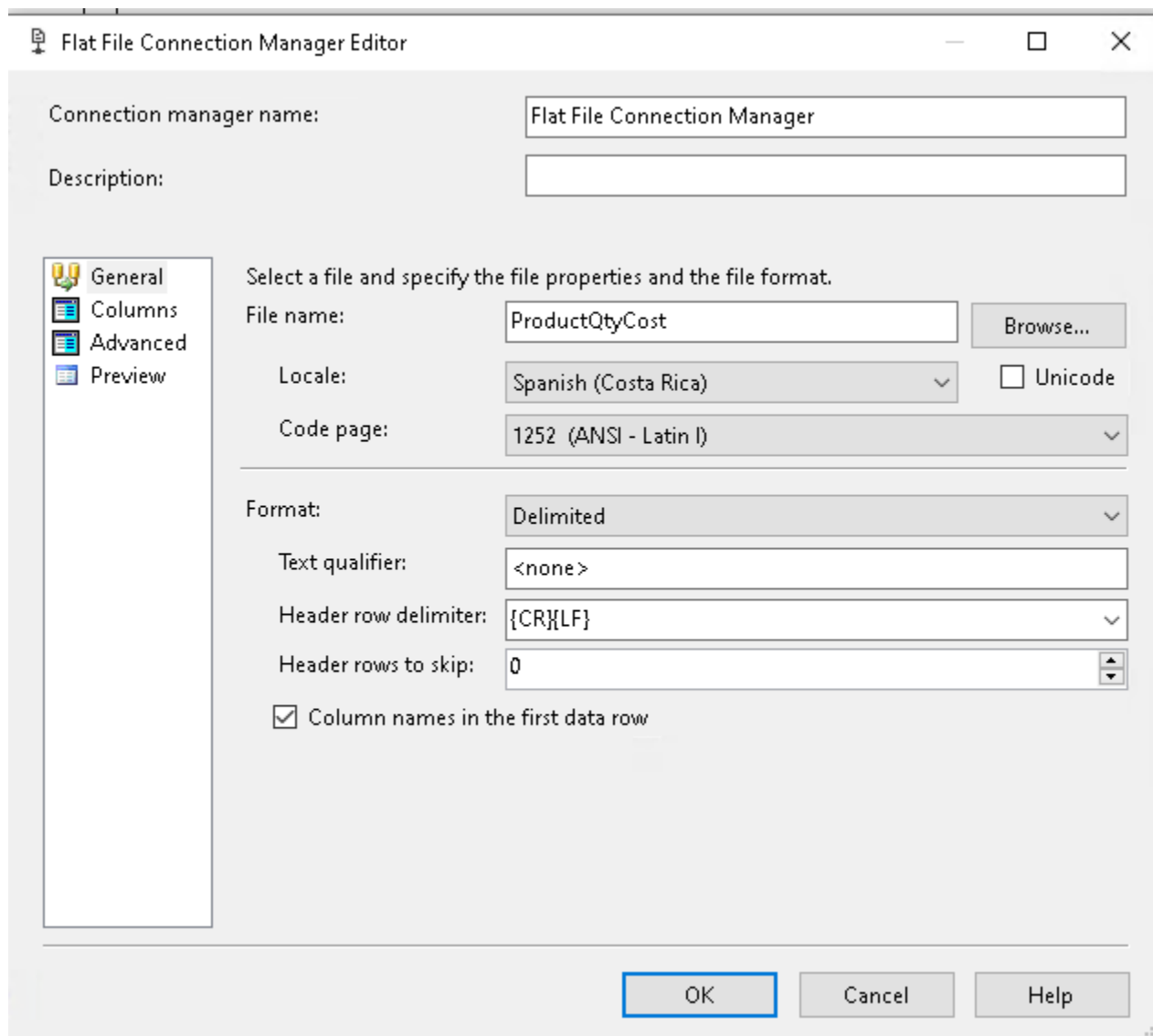


Figura 4. El editor de conexión de Flat File Destination.

En el editor de *Flat File* seleccionar la opción *Column* y asegurarse que el mapeo entre columnas esté correcto como se muestra en la figura 5.

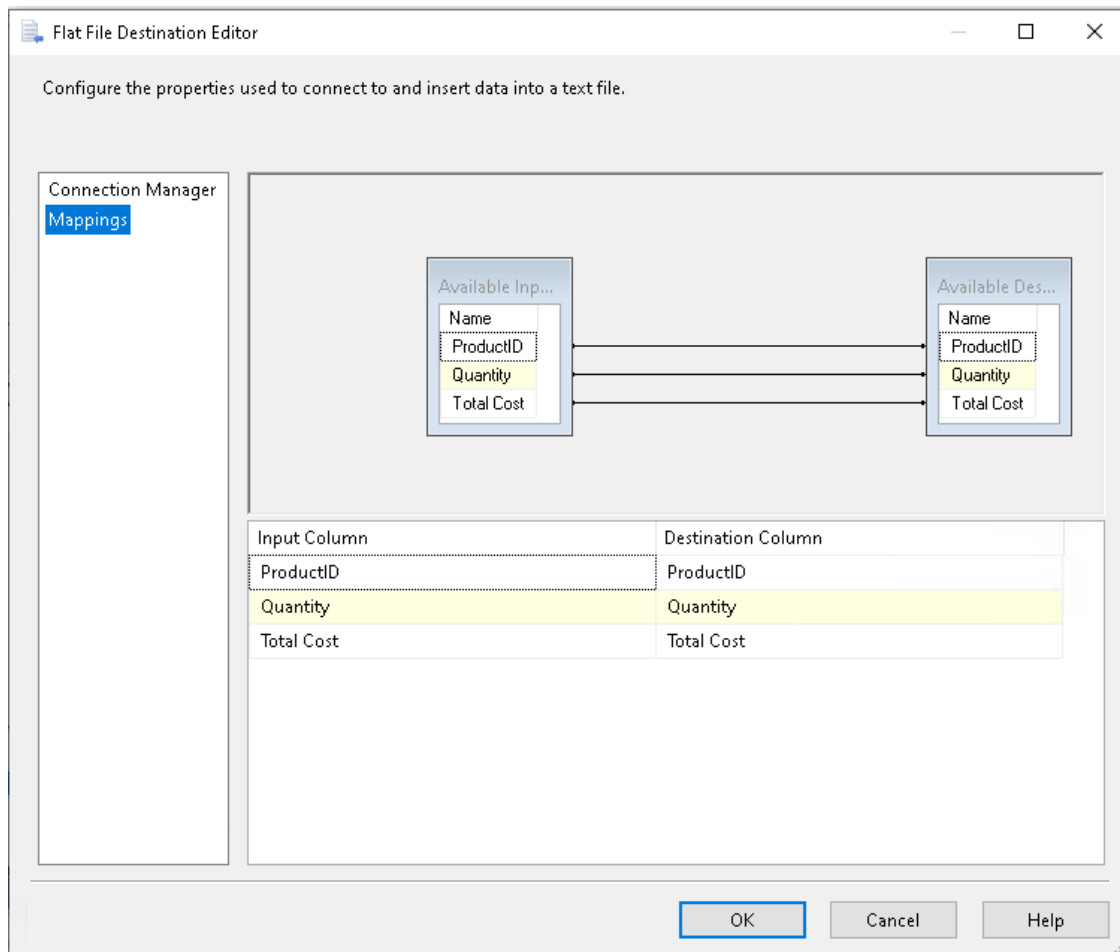


Figura 5. Editor de destinación de Flat File Destination

IV. Ejecución del paquete

El paquete debe tener las componentes como se muestra en la figura 6.

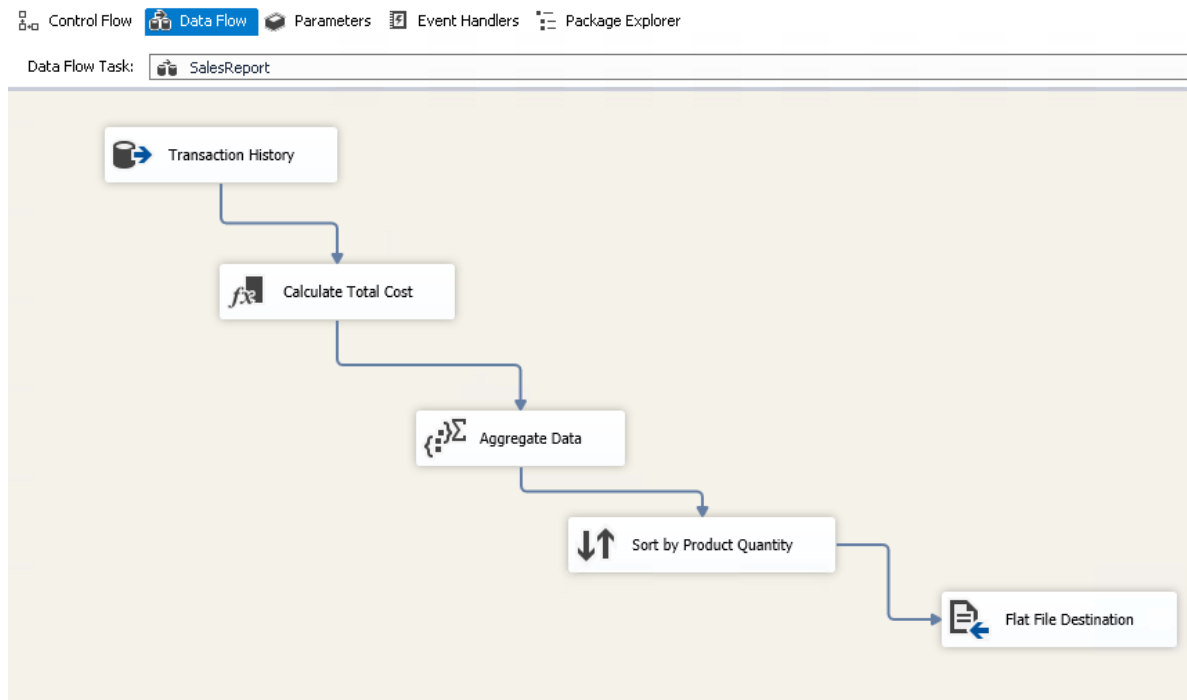


Figura 6. Componentes del paquete

1. Ejecutar el paquete y observar en la pestaña *Progress* los pasos que se realizan. Para volver al modo de diseño de paquete se necesita parar el proceso de *debugger*.
2. Revisar el contenido del archivo de destino.

V. Vista de los datos durante la ejecución del paquete

1. Para poner las vistas de datos en el flujo de datos, con el clic derecho en la flecha azul saliendo del origen de datos o de alguna transformación seleccionar *Data Viewers*.
2. Seleccionar el tipo y las columnas que desea ver.
3. Ejecutar el paquete. La ejecución se detiene para enseñar los datos requeridos. Se necesita presionar el botón de > o cerrar el *Data Viewer* para proseguir la ejecución.
4. Realizar las pruebas con dos diferentes *Data Viewers*.

VI. Envío simultaneo a diferentes destinos

A veces puede ser necesario mandar los mismos datos a diferentes formatos de archivos, por ejemplo, Excel, archivo plano.

1. Modificar el paquete borrando la conexión entre “Aggregate Data” y “Sort by Product Quantity”.
2. Arrastrar la transformación de *Multicast* y ponerle el nombre “Multiple Destinations”.

3. Conectar “Aggregate Data” con “Multiple Destinations” y “Multiple Destinations” con el anteriormente creado “Sort by Product Quantity”.
4. En la forma parecida establecer el ordenamiento por Product ID y enviar los resultados a la hoja Excel (debe primero crearla) usando otro camino de flujo de datos que sale de “Multiple Destinations”.

Nota: al crear una nueva tabla de Excel, puede ser necesario modificar el tipo de datos por INT para ProductID y NUMERIC (10,2) para los demás atributos y además, modificar la configuración como se muestra en la figura 7 (Project Properties -> Debugging -> Run64BitRunTime = False) para correr en ambiente de 32 bits. (Documente todos los cambios que realice y las fuentes consultadas.)

5. Ejecutar el nuevo paquete.
6. Verificar los datos en la hoja Excel en una máquina cliente que tenga instalado el software correspondiente.

**PARA HACER CUALQUIER COSA DEBEMOS PRIMERO CREER
QUE PUEDE HACERSE;
EL HECHO DE CREER QUE ALGO PUEDE LOGRARSE
PONE EN MOVIMIENTO LA MENTE
PARA ENCONTRAR LA MANERA DE HACERLO**

David J. Schwartz