MANUAL

DATOS ABIERTOS

EN COLOMBIA

PARA:

CREACIÓN DE UN ETL

En el siguiente documento se explica el procedimiento para realizar las diferentes acciones que posibilita la herramienta Socrata

MINTIC

ASESOFTWARE 25 AÑOS

vive digital para la gente

TODOS POR UN NUEVO PAÍS
PAZ EQUIDAD EDUCACIÓN

TABLA DE CONTENIDOS

01 PROPÓSITO DEL DOCUMENTO

En el presente documento se pretende recoger información de utilidad, a modo de tutorial, para desarrolladores quienes tengan rol de editor, publicador o administración en la plataforma de datos abiertos para el gobierno colombiano, y quienes están enfocados en un desarrollo de un ETL, con el fin de automatizar la carga de datos abiertos al portal Socrata.

02 REQUISITOS PARA LA CREACIÓN DE UN ETL

En el siguiente listado aparecen los requisitos básicos para crear un ETL, en este tutorial veremos cómo crearla a partir de SQL Server Business Inteligences y DataSync, pero se puede usar cualquier herramienta de ETL, tal como aparece en la tabla a continuación:

Herramienta	Complejidad	Programable	Transformación	Costo
Manual	★☆☆	✗	✗	
SoDA API	★★★	✓	✓	$$$
DataSync	★★☆	✓	✗	
Safe FME	★★☆	✓	✓	$
ETL (Pentaho Kettle)	★★★	✓	✓	$$

- Visual Studio 2008 o última versión.
- SQL Server 2008 o última versión.
- Integration services.
- Business Inteligences.
- JAVA 7 JDK o última versión.
- Descargar dataSync de https://github.com/socrata/datasync/releases

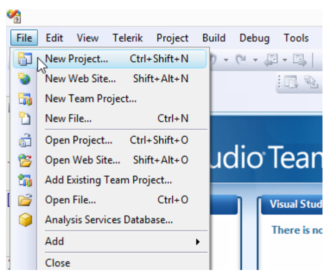
03

PASO A PASO

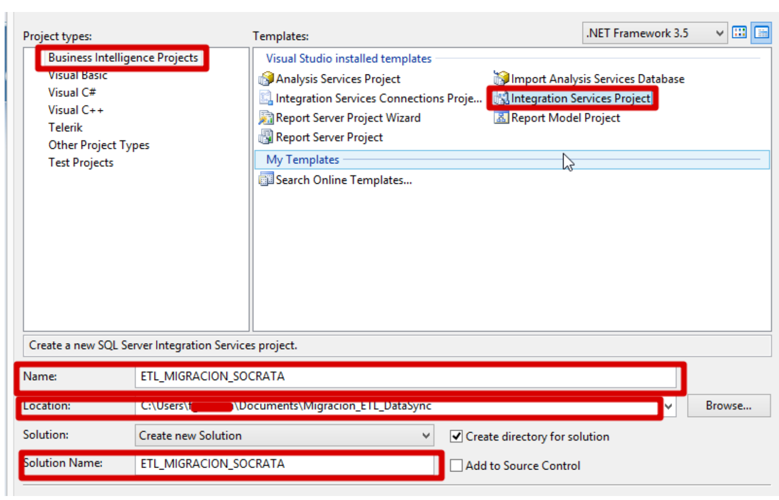
3.1. CREAR UN ETL

A continuación, listaremos paso a paso, como crear una sencilla ETL para automatizar la carga de datos abiertos a Socrata, es indispensable contar con un ambiente de Base de Datos si se quiere seguir este tutorial.

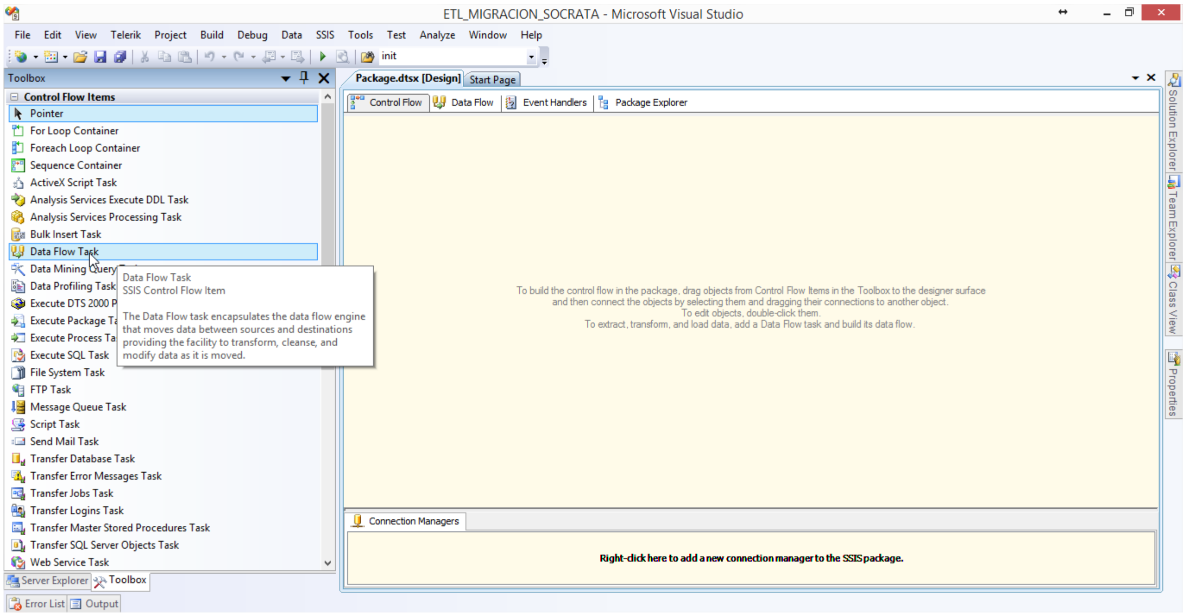
I. Creamos un nuevo proyecto de ETL en nuestra herramienta de visual estudio.



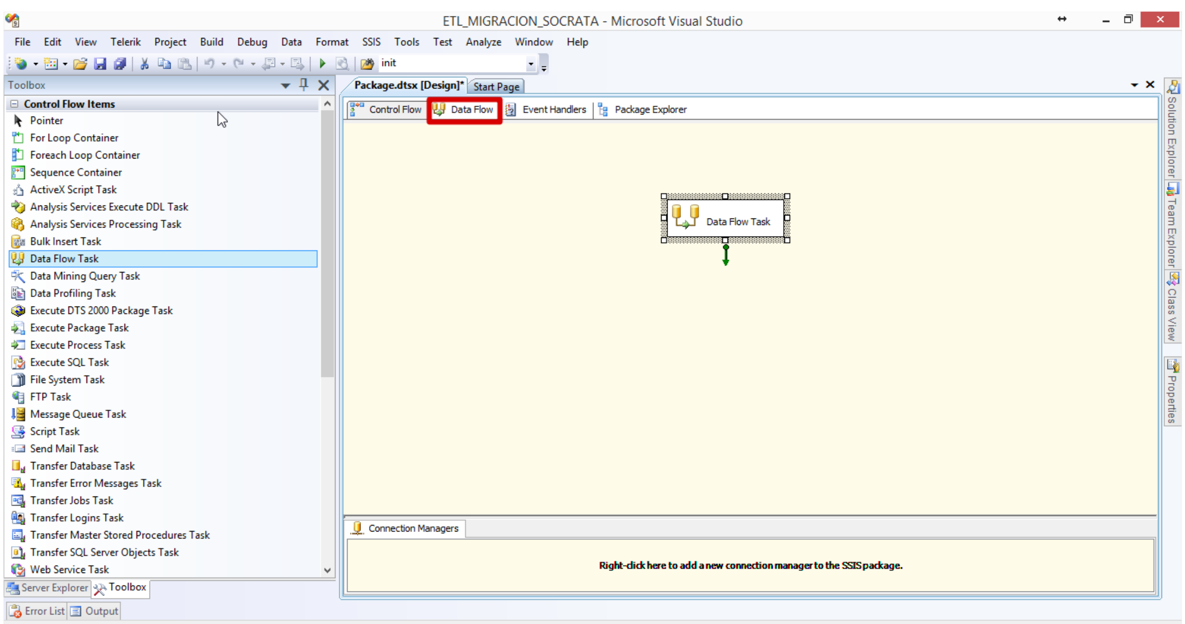
II. Seleccionamos Business *Intelligence Projects – Integration Services Project*, es necesario colocar un nombre al proyecto y especificar la ruta respectiva donde queremos alojar nuestro proyecto.



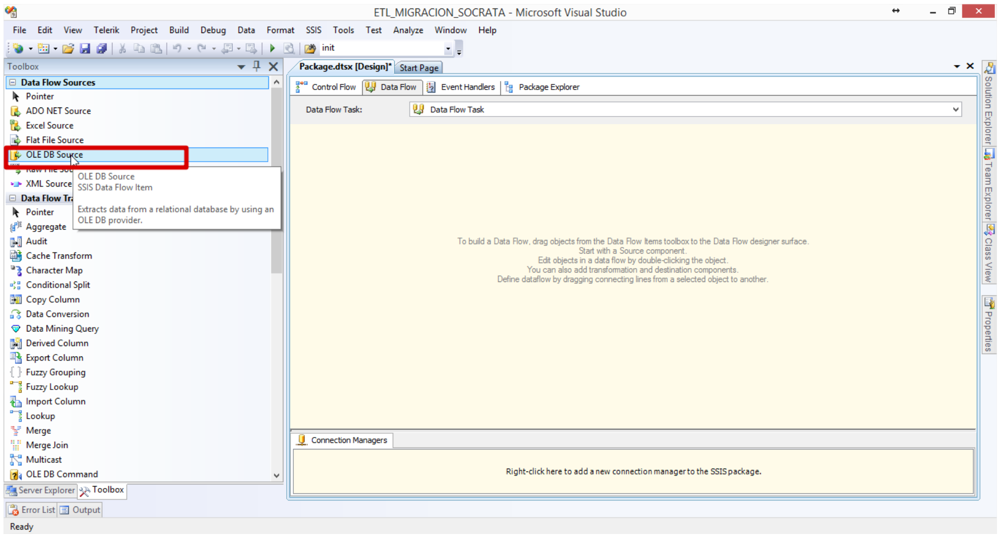
III. En nuestro nuevo proyecto, seleccionamos el componente *Data Flow Task*. Arrastramos este componente hacia la ventana de *Control Flow*.



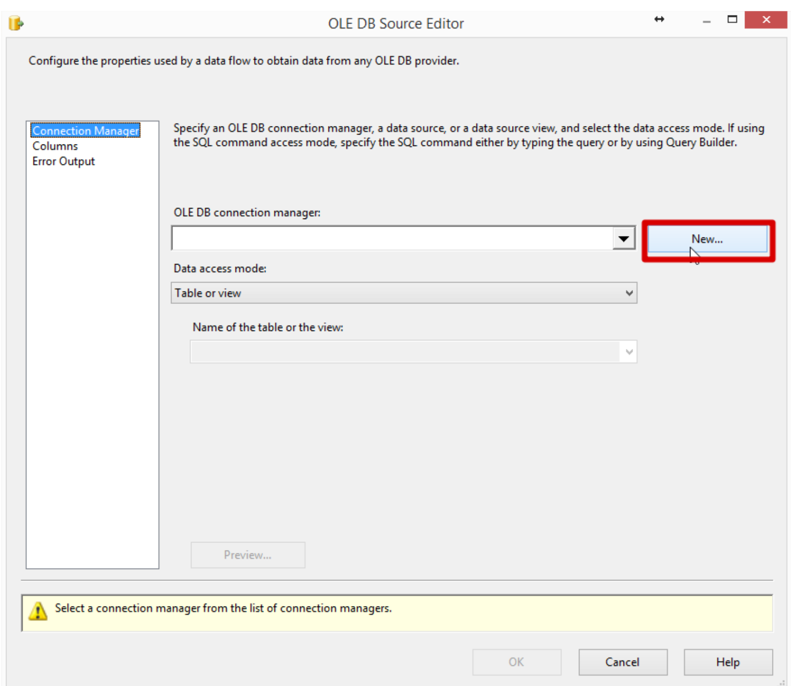
IV. En la parte superior de la venta seleccionamos la pestaña *Data Flow* o también podemos dar doble clic sobre el componente mismo.



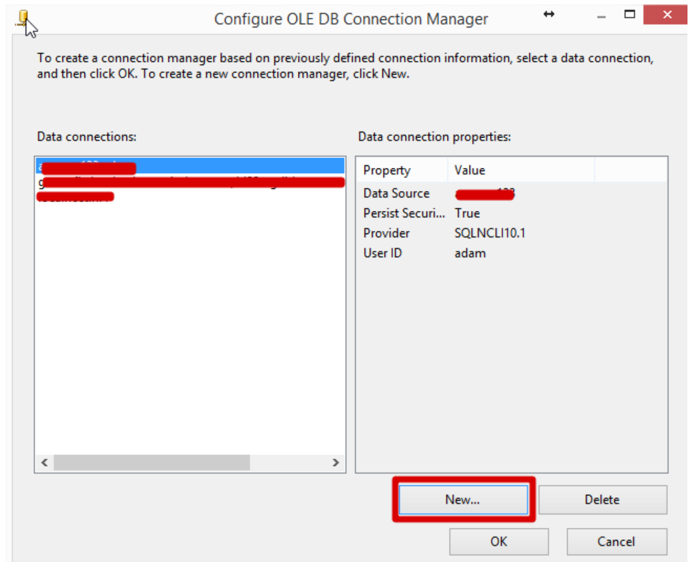
V. En nuestro *Data Flow*, seleccionamos el componente *OLE DB source* y lo arrastramos a nuestra ventana *Data Flow*.



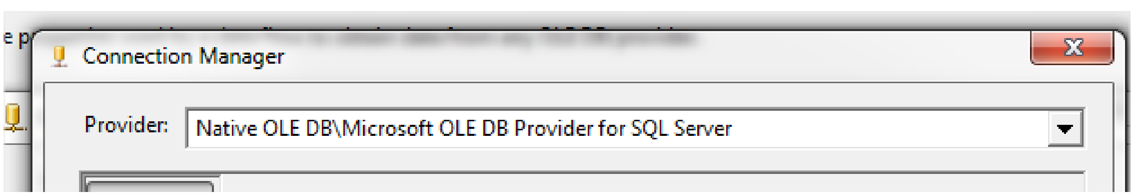
VI. Al darle doble clic al componente nos despliega la ventana de configuración, seleccionamos el botón "*New…*"

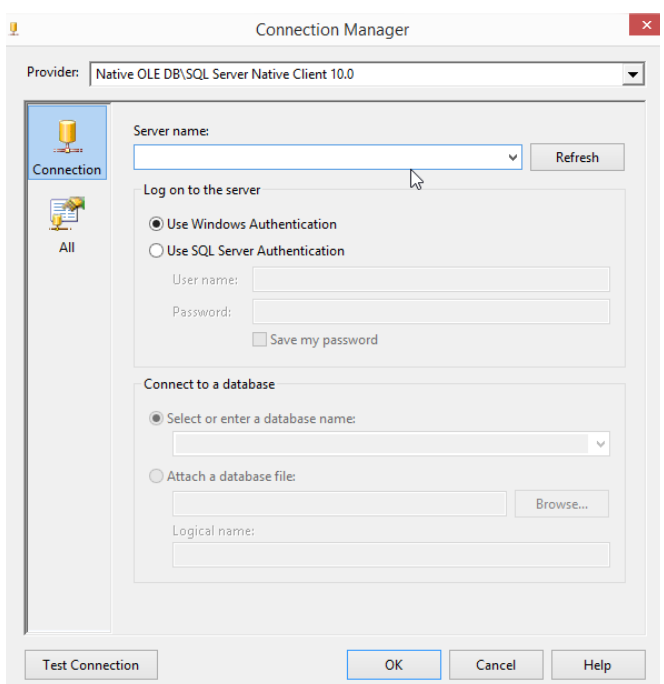


VII. Creamos nuestra nueva conexión OLE DB dando clic nuevamente en “*New…*”

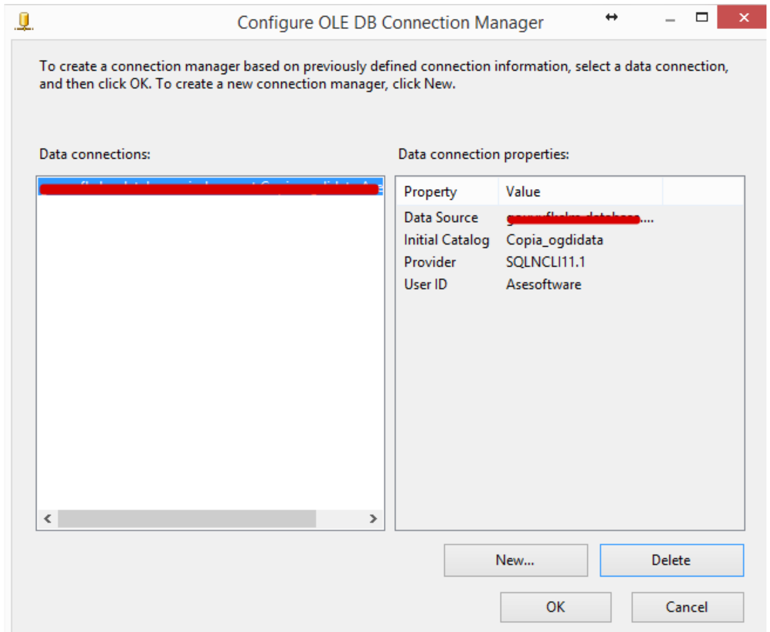


VIII. Diligenciamos todos nuestros datos de conexión y al finalizar damos clic en “*OK*”.

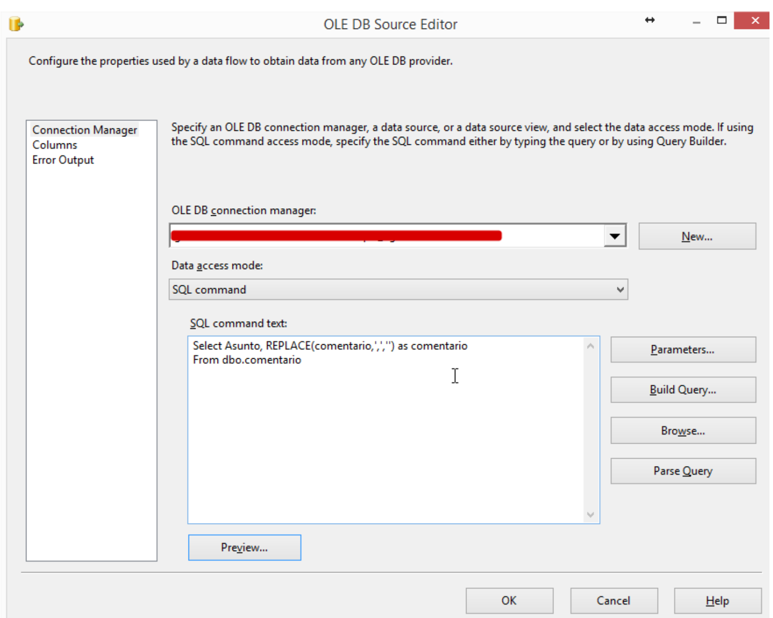




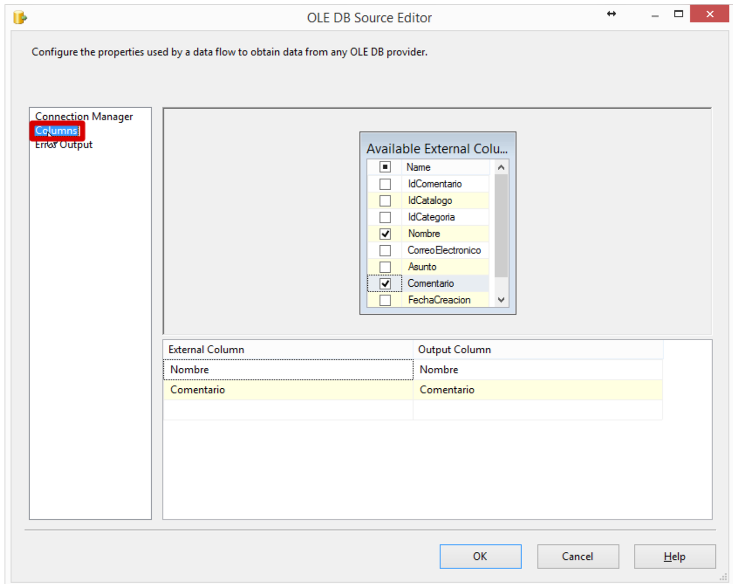
IX. Nos mostrara en pantalla la nueva conexión, debemos seleccionarla y luego damos clic en “*OK*”.



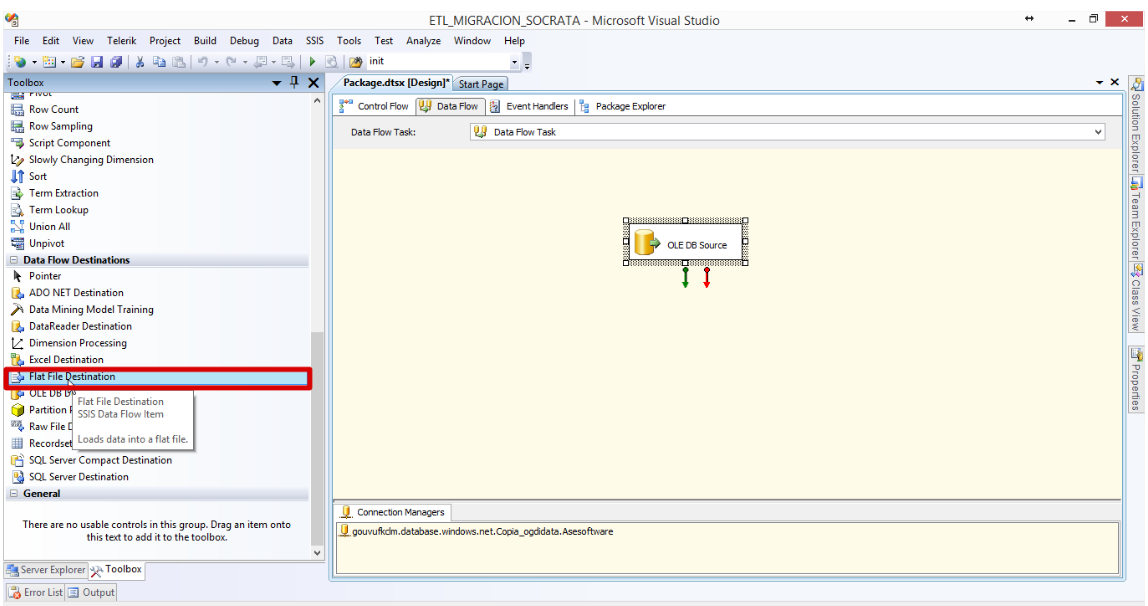
X. Volvemos a nuestra pestaña de *OLE DB Source* donde seleccionamos “*SQL command*” en “*data acces mode*”, allí especificaremos nuestra consulta de datos al finalizar podemos seleccionar *preview* para revisar que la consulta sea correcta.



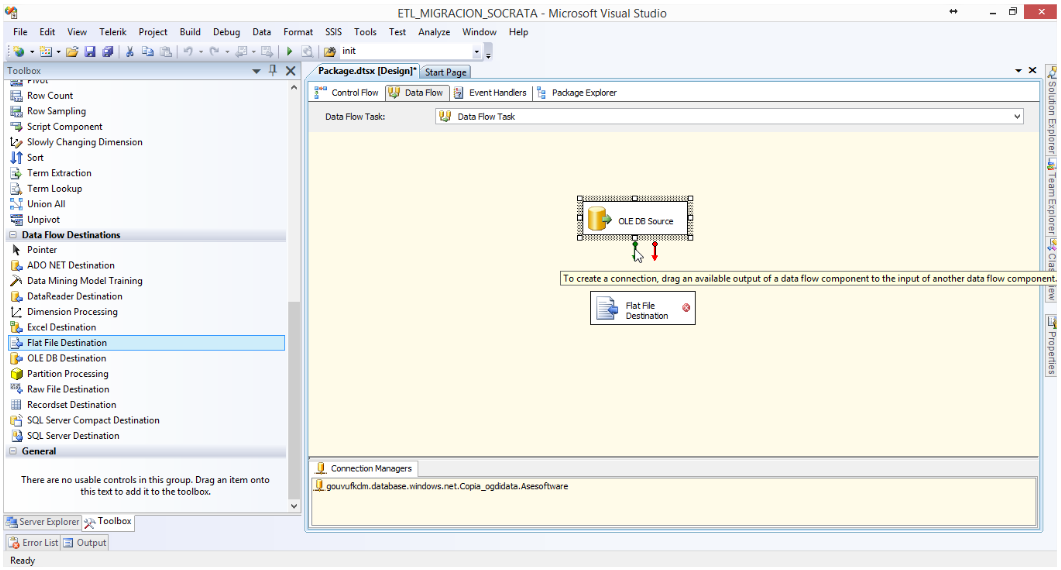
XI. En la parte superior derecha seleccionamos la opción “*Columna”*, donde podemos seleccionar las columnas que queremos enviar de la consulta especificada en el ítem anterior. Damos clic en “*OK*”.



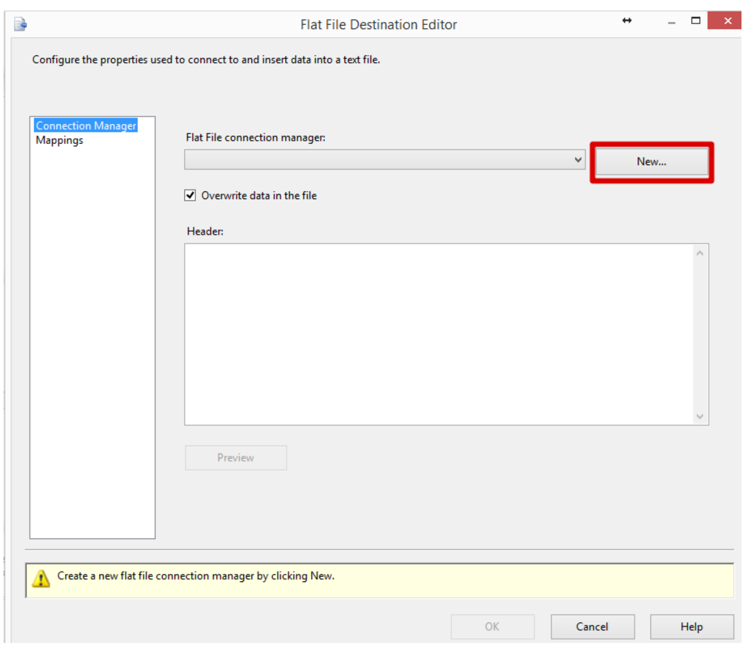
XII. Ahora seleccionamos el componente *File Flat Destination* y lo arrastramos en nuestra ventana, *Data Flow*.



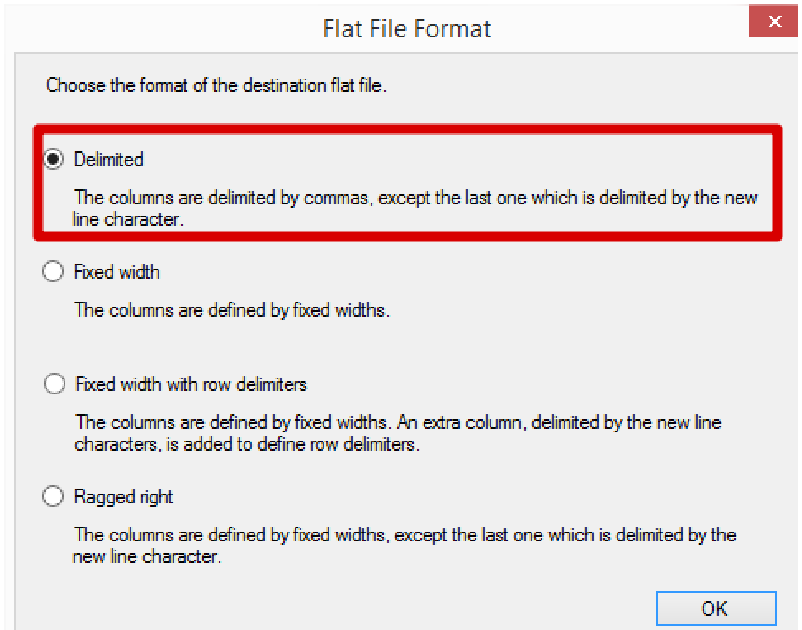
XIII. Unimos nuestros dos componentes por medio de la flecha verde que indica que las tareas deben ser ejecutadas en secuencia y abrimos nuestro componente *File Flat Destination*.



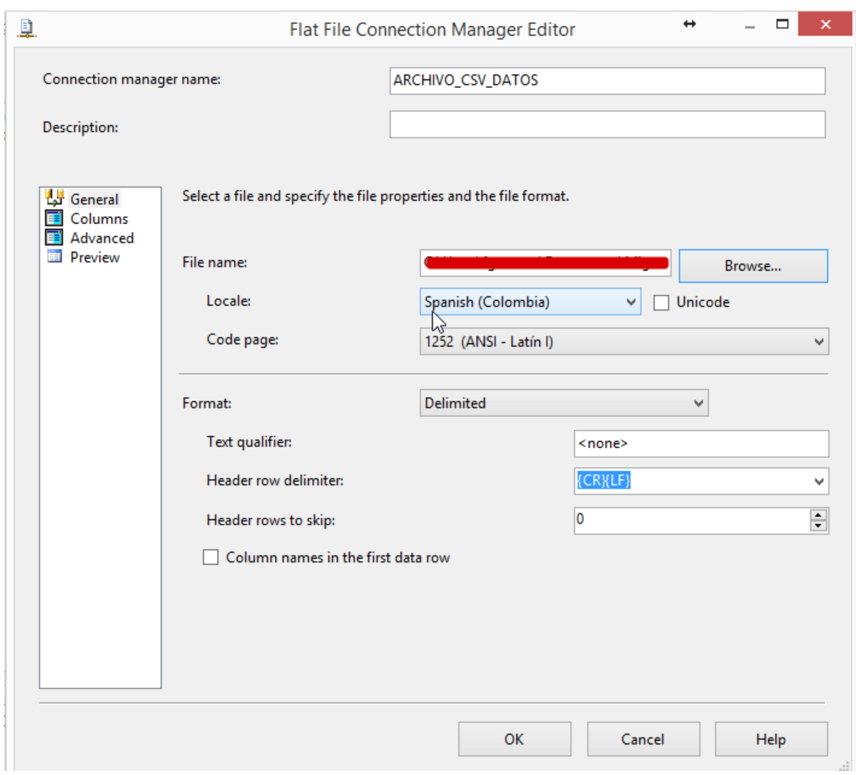
XIV. En la ventana de configuración seleccionamos "*New…*" donde se creará el archivo CSV donde guardaremos los datos resultados de nuestra consulta de base de datos.



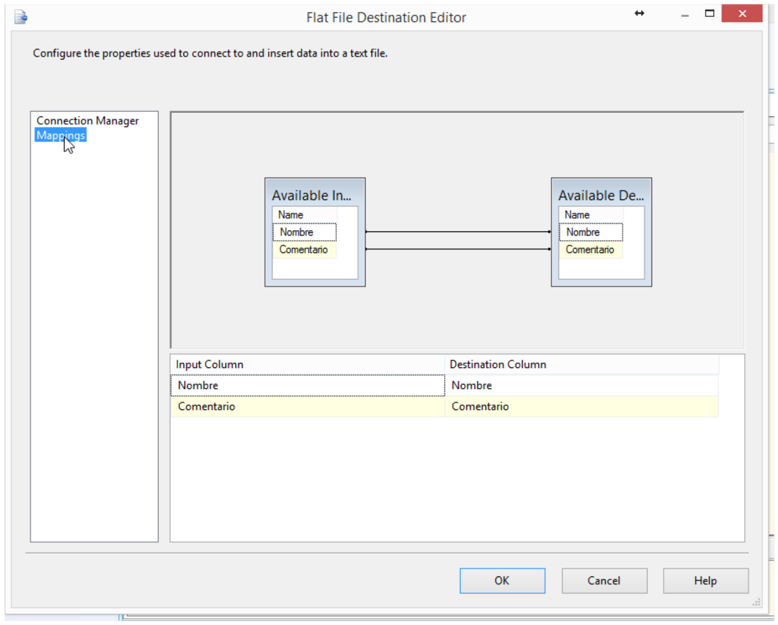
XV. Se selecciona el formato por el cual se va a separar el archivo, el cual va a ser Delimitado por comas, damos clic en “*OK*”.



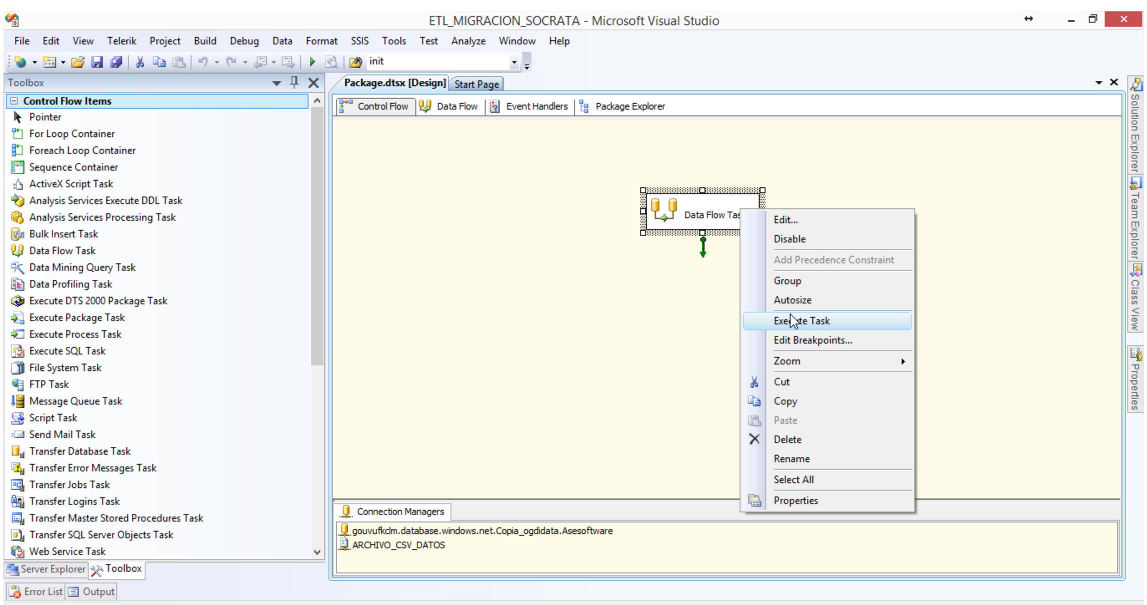
XVI. Ahora debemos especificar el nombre del *flat file* y seleccionamos la ubicación donde vamos a generar nuestro archivo, al finalizar clic en “*OK*”.



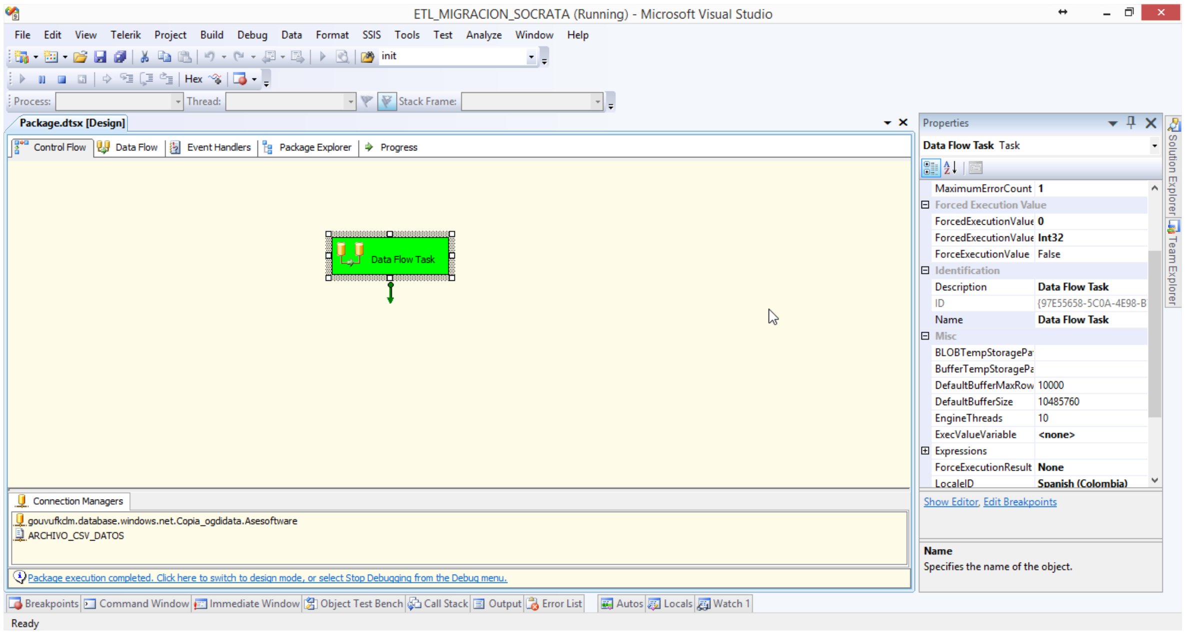
XVII. En la pestaña de *Mappings* seleccionamos la columna de origen y destino, para posteriormente unirlas, al finalizar clic en “*OK*”.



XVIII. Retornando a la ventana del *control flow*. Ejecutamos nuestro proyecto seleccionando con el ratón el clic derecho y posteriormente “*Execute Task*”



XIX. El proyecto se ejecuta exitosamente, podemos comprobarlo observando el nuevo archivo que se generó en la ruta que especificamos.



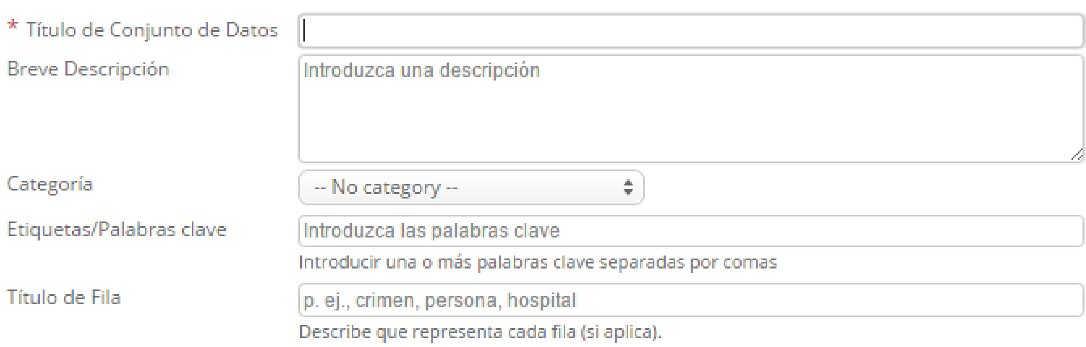
3.2. ESTRUCTURA EN SOCRATA

Es indispensable para la ETL contar con una estructura en Socrata del conjunto de datos, es decir, una tabla con las columnas que se van a cargar desde la ETL.

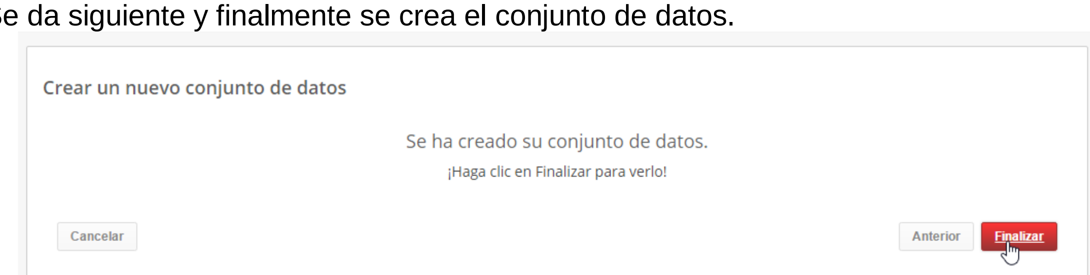
I. Se debe ingresar a la página de datos abierto https://www.datos.gov.co/, iniciar sesión con el usuario con rol de publicador, editor o administrador, y crear un nuevo conjunto de datos.

II. Seleccionar *diseñar de cero*, y clic en siguiente.

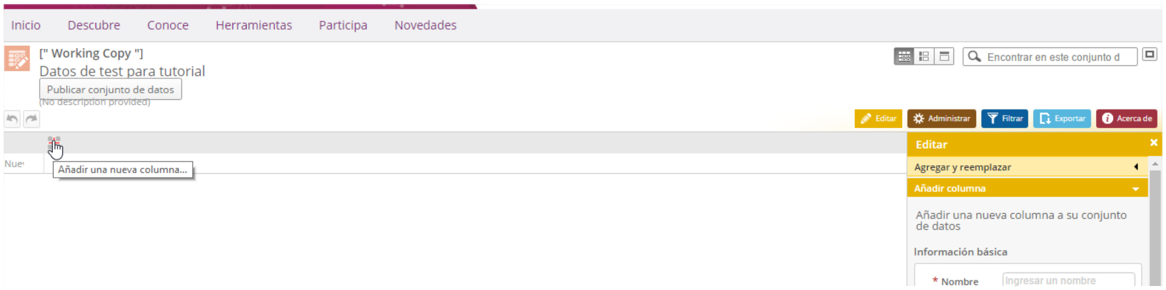
III. A continuación, debemos diligenciar la página de la metadata del conjunto de datos correctamente, si se tiene dudas sobre esto se puede ver el manual de *publicador sección Diligenciamiento de los Metadatos.*



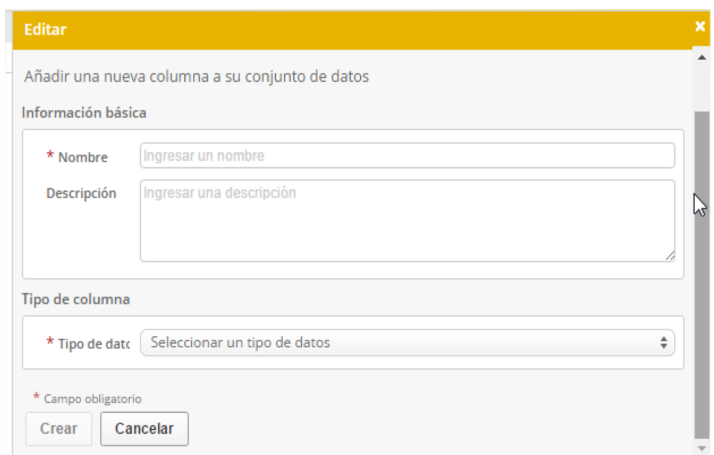
IV. Se da siguiente y finalmente se crea el conjunto de datos.



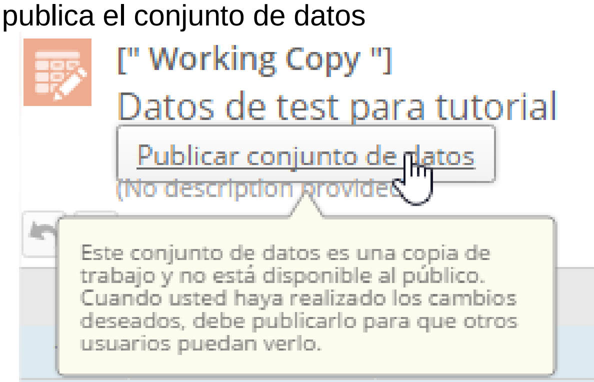
V. Luego de creada, se añade los campos que concuerden con los datos que se deseen exportar, es decir deben ser los campos resultados de nuestra consulta de base de datos del ETL en el numeral anterior.



VI. Al añadir un campo, se debe nombrar y colocar el tipo de dato adecuado, según la data exportada en la ETL, para ello en la parte derecha de la pantalla se despliega la ventana "*Editar*" para la especificación de estos datos.



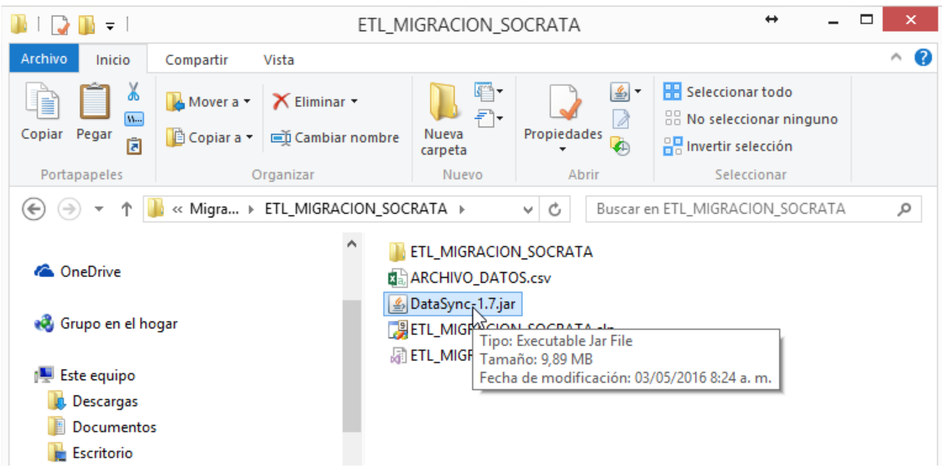
VII. Para finalizar se publica el conjunto de datos



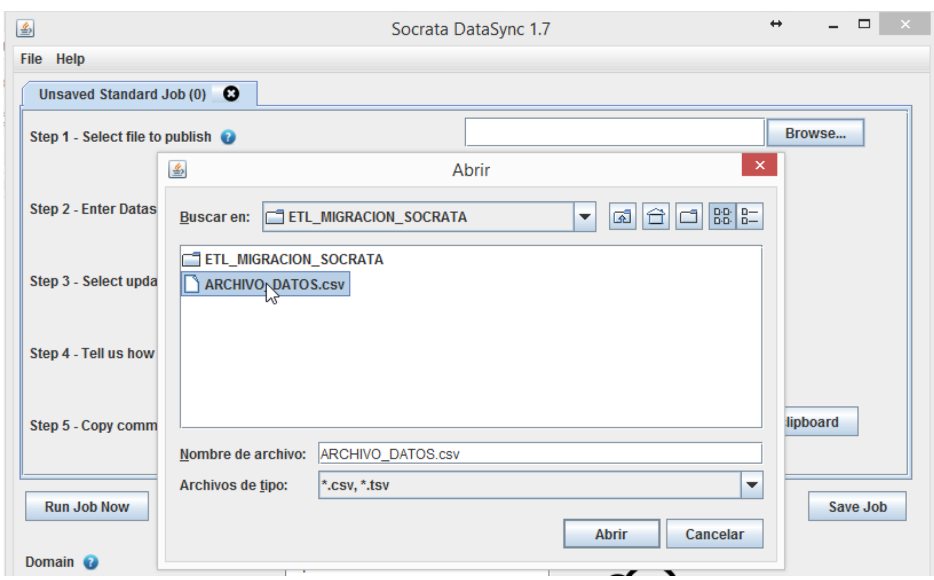
3.3. CONFIGURANDO DATASYNC

Nuestro último paso es configurar el DataSync a partir del resultado del ETL y utilizando la estructura creada en Socrata.

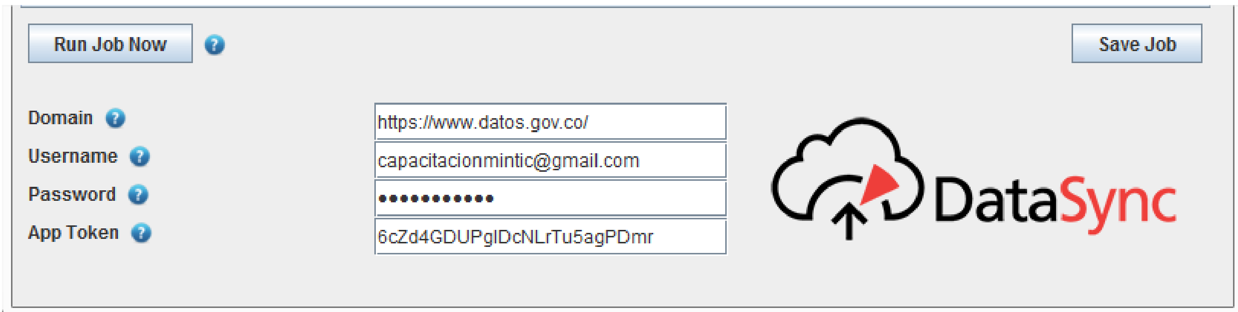
I. Debemos dirigirnos a la ruta donde se encuentra el ejecutable DataSync (Última versión disponible en https://socrata.github.io/datasync/), el cual desplegamos, si el DataSync no despliega correctamente es necesario volver a la sección dos de este tutorial y comprobar que todos los requisitos, están correctamente instalados.



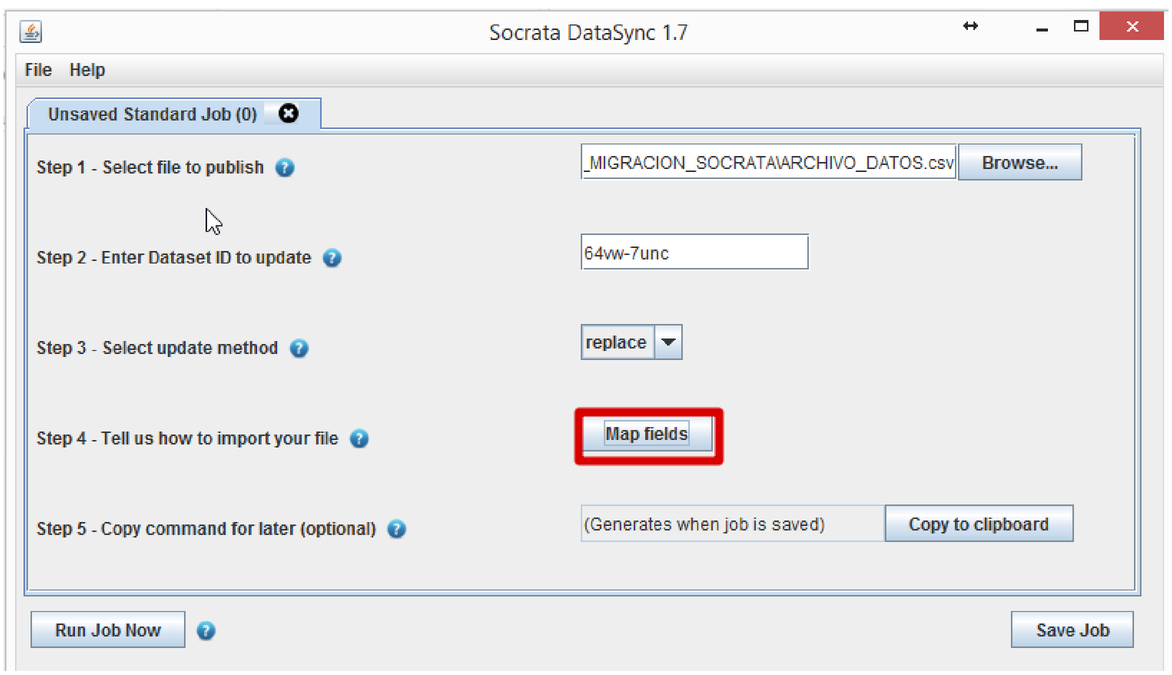
II. Debemos seleccionar nuestro archivo, generado anteriormente con la ETL.



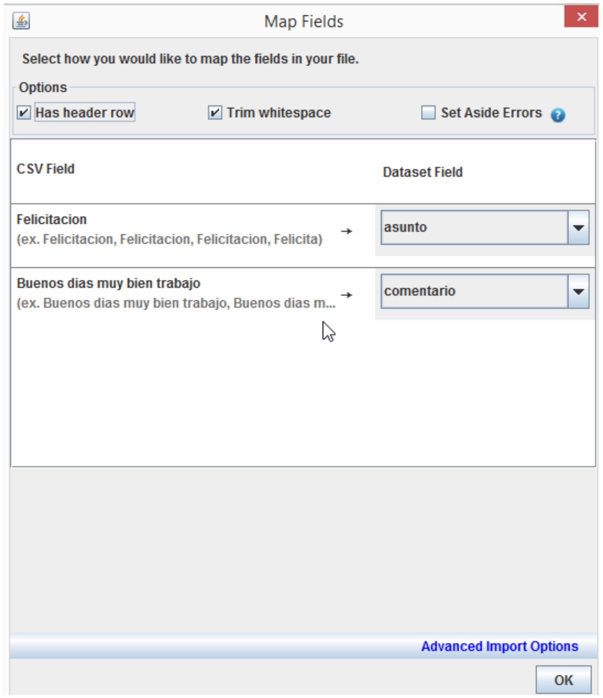
III. En la parte inferior completamos los datos de nuestro usuario y con su respectivo A*pp Token* creado anteriormente (ver Manual de desarrollador sección 4 *DataSync,* para más información).



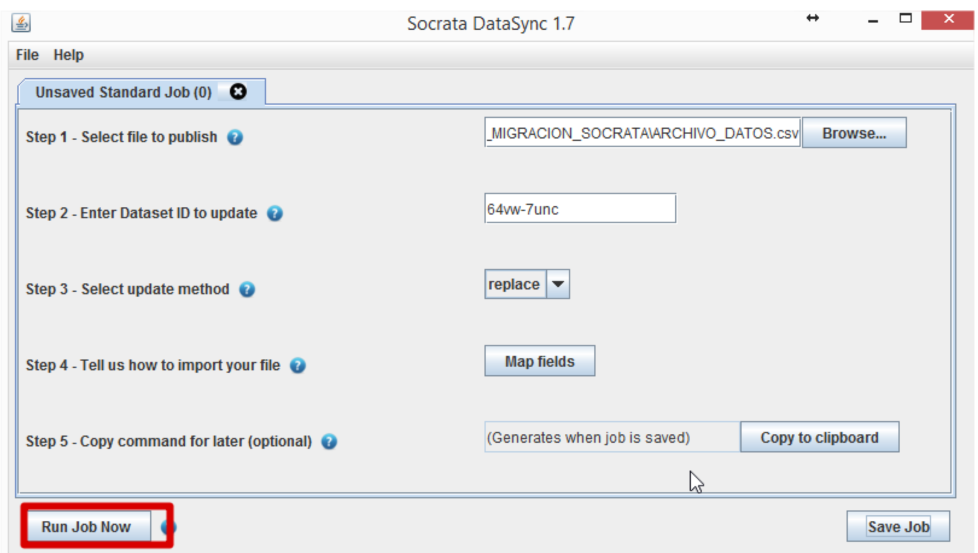
IV. Posteriormente seleccionamos el ID del dataset donde queremos migrar nuestros datos (ver seccion anterior de este manual). Seleccionamos el metodo de actualización y los campos que deseamos mapear.



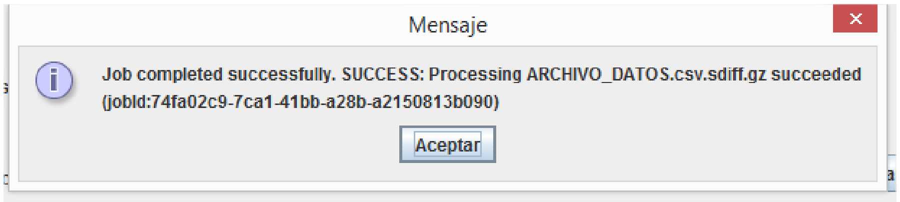
V. En *map fields* seleccionamos los campos que vamos a mapear.



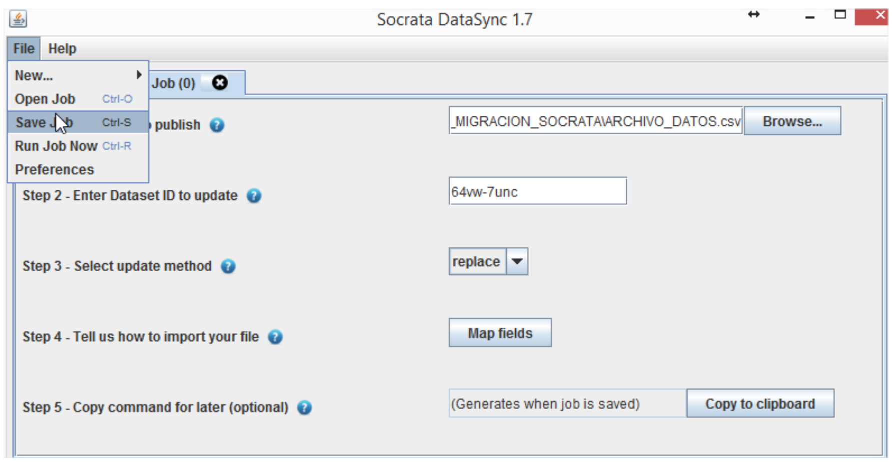
VI. Ejecutamos el job y certificamos que este funcione correctamente



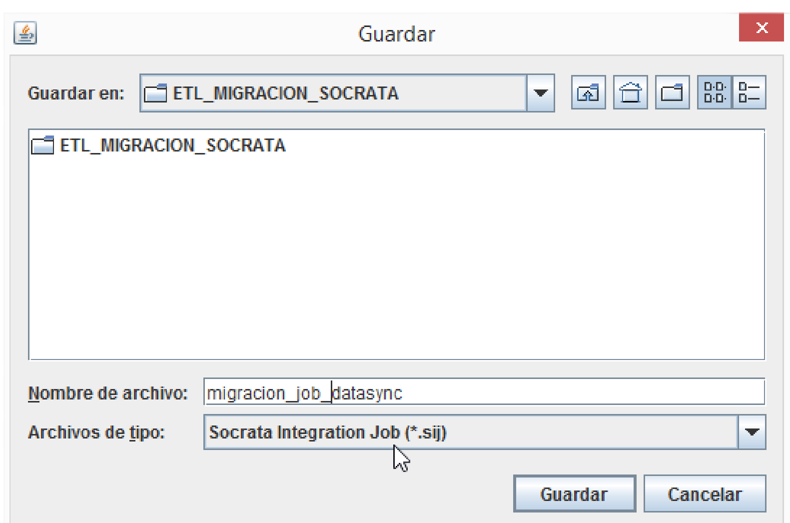
VII. Si el job se ejecuta exitosamente, mostrara un mensaje como el siguiente.



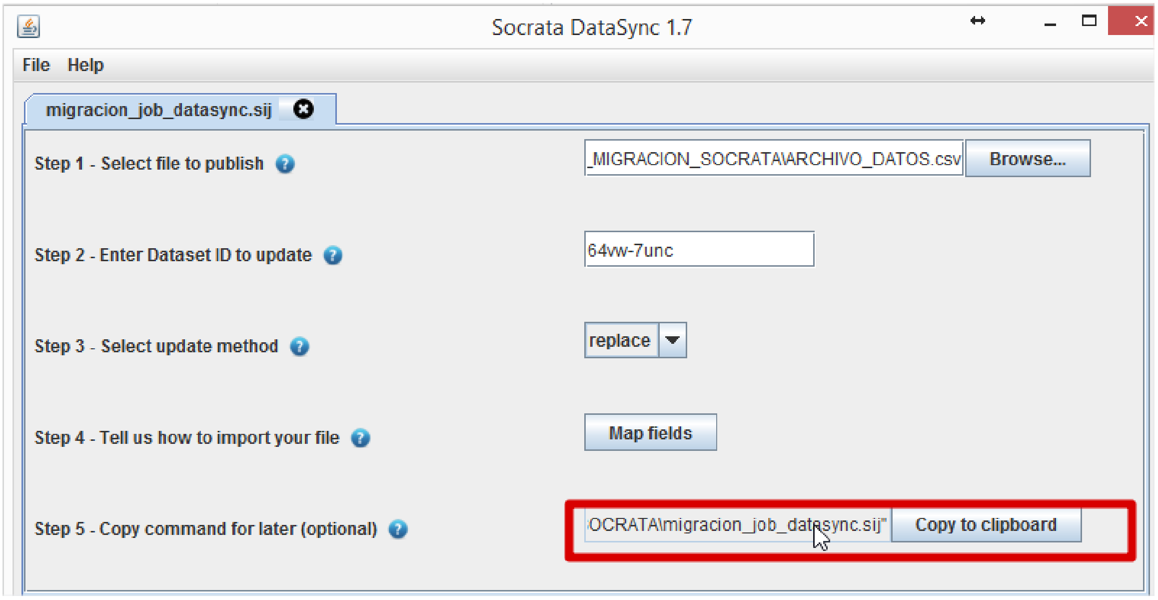
VIII. Debemos guardar el job especificando una ruta.



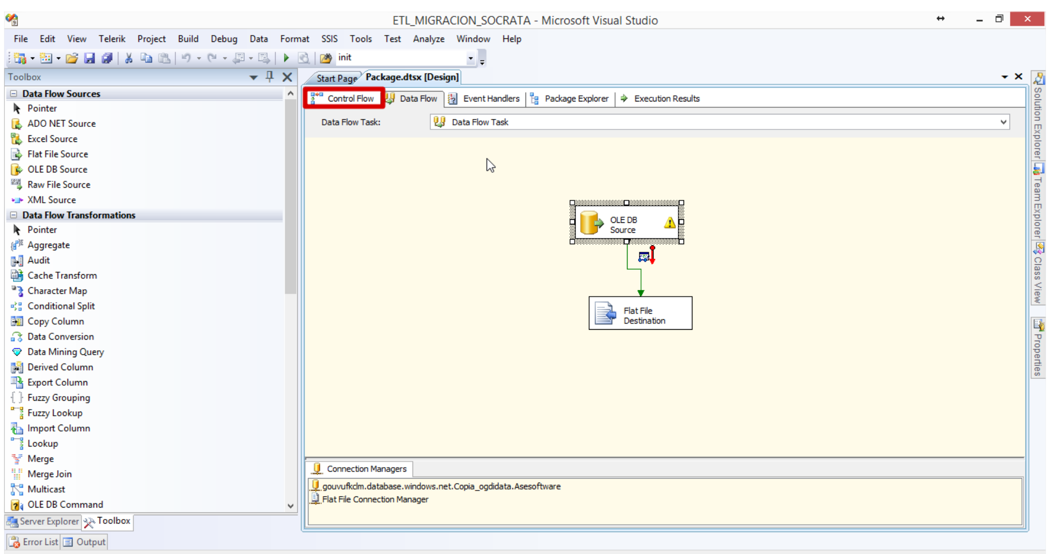
IX. Colocando un nombre adecuado para el job.



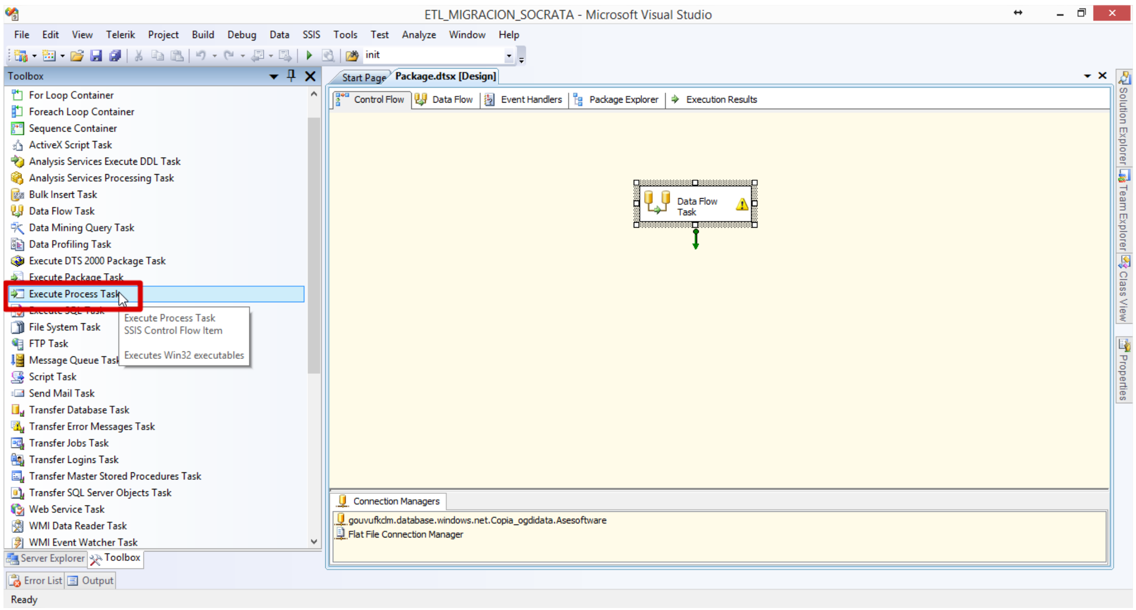
X. Al guardar el campo del comando se completa, este campo lo vamos a utilizar para nuestra ETL en los pasos posteriores.



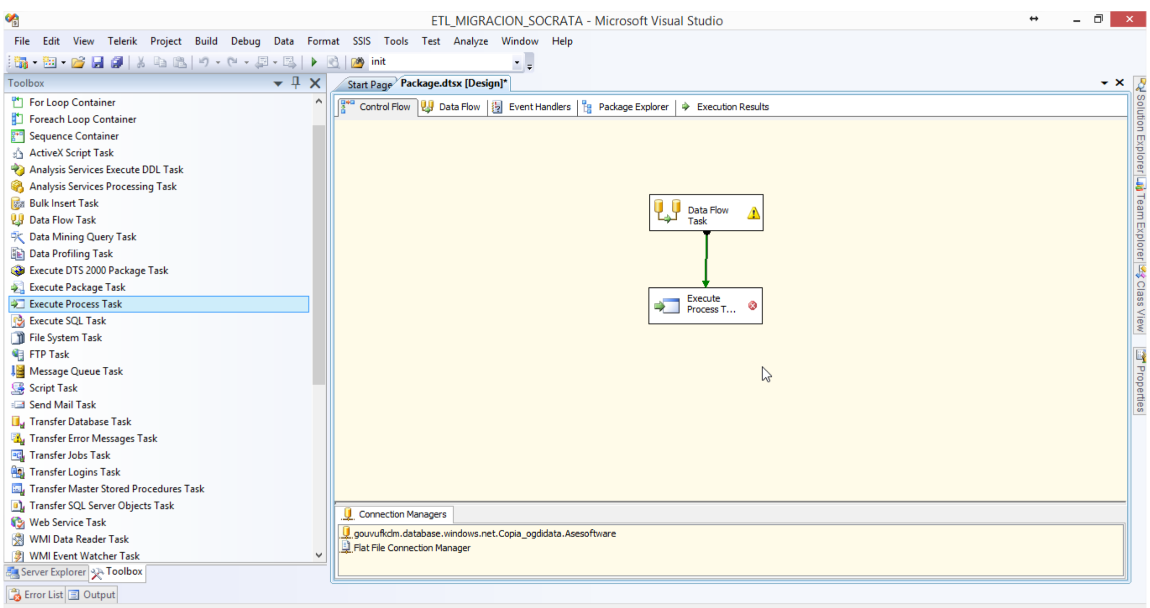
XI. Volvemos a nuestra ETL y seleccionamos la pestaña *Control Flow*.



XII. Seleccionamos el componente *Execute Process Task* y lo arrastramos a la ventana.

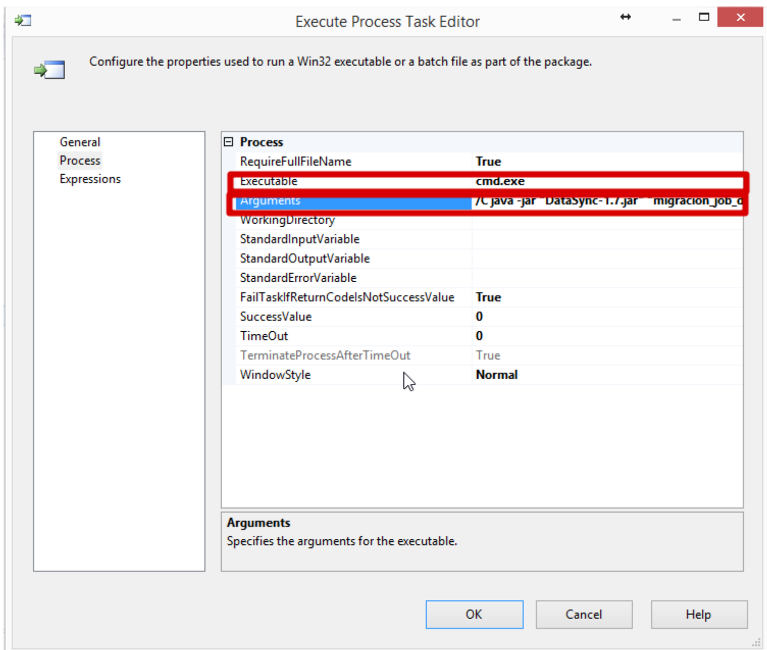


XIII. Unimos estos dos componentes mediante la flecha verde.

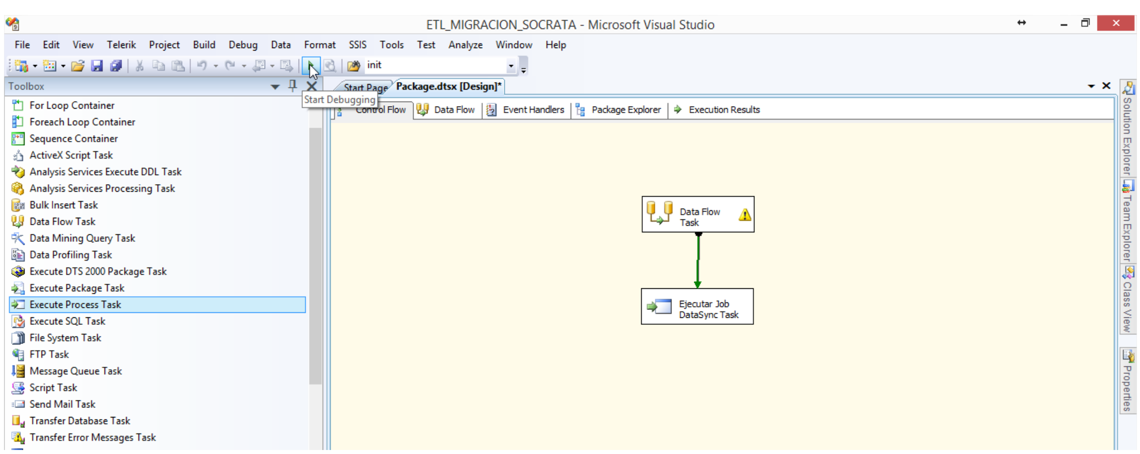


XIV. En *Executable* se ingresa "cmd.exe", en el campo *Arguments* se ingresa el comando que nos genera la herramienta Datasync agregando el comando "/C" así:

/C java –jar "ruta donde se encuentra el programa DataSync-1.7.jar" "ruta donde se guardó el job de datasync"



XV. Al finalizar clic en "OK"



XVI. Por ultimo ejecutamos nuestra ETL. Ejecutándose exitosamente, notaremos que nuestro conjunto de datos en Socrata ha sido poblado con la consulta que indicamos en la ETL.