TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

---------- ----------

BÁO CÁO ETL.Assigment

Môn: Hệ thống thông tin phục vụ trí tuệ kinh doanh

7/11/2024 – 13/11/2024

MÃ HỌC PHẦN: CSC12107

ID	**Họ tên**
21127699	Lô Thủy Tiên

Giảng viên:
ThS. Hồ Thị Hoàng Vy
ThS. Tiết Gia Hồng
ThS. Nguyễn Ngọc Minh Châu

Báo cáo:
Lô Thuỷ Tiên

Thành phố Hồ Chí Minh – 2024

MỤC LỤC

THÔNG TIN VỀ ĐỒ ÁN

Mã học phần: CSC12107

Tên học phần: Hệ thống thông tin phục vụ trí tuệ kinh doanh

Tên : BÁO CÁO ETL.Assigment

Hình thức:

- Báo cáo (.doc, ppt)

- Source demo/ Video demo

- MSSV_SSIS

Mô tả:

Project SSIS thực hiện ETL với các bước sau:

1. **Extract dữ liệu từ Excel** (Sheet `Học sinh` và `LopHoc`) và đưa vào **Stage** theo phương pháp **Incremental Extract**.
2. **Load vào Stage**: Làm sạch, loại bỏ dữ liệu cũ, và cập nhật dữ liệu mới.
3. **Load từ Stage sang NDS** (`LopHoc_NDS`, `HocSinh`) sau khi đã profiling và làm sạch.
4. **Metadata** ghi lại thời gian ETL (LSET, CET) để kiểm soát dữ liệu rút trích.
-

Giảng viên phụ trách: Cô Hồ Thị Hoàng Vy, Cô Tiết Gia Hồng, Cô Nguyễn Ngọc Minh Châu

I. MÔ TẢ DỮ LIỆU NGUỒN (SOURCE):

1. Mô tả dữ liệu:

Sheet LopHoc (chứa thông tin về lớp học):

Tên thuộc tính	Ý nghĩa
MaLop	Mã lớp, là định danh duy nhất cho mỗi lớp học
TenLop	Mã lớp, là định danh duy nhất cho mỗi lớp học

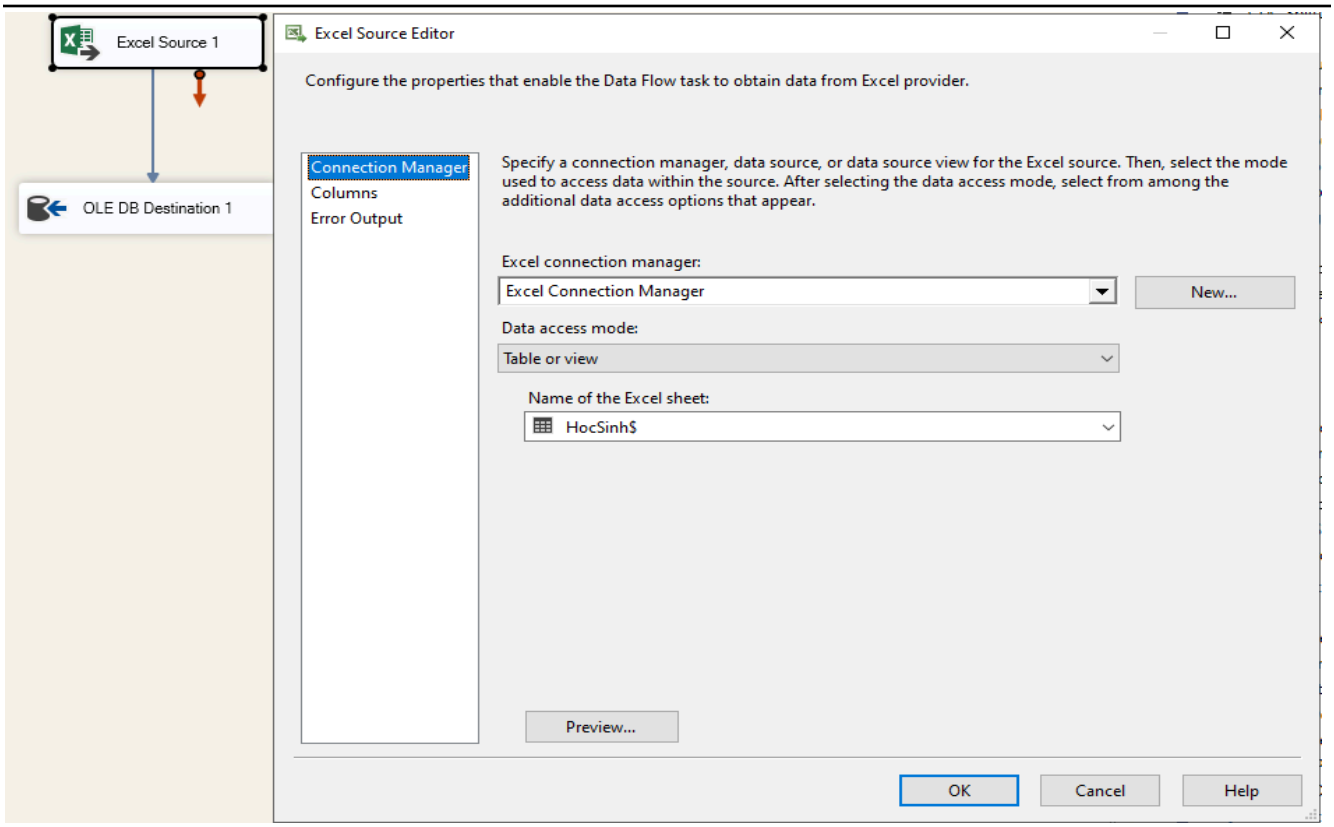
Sheet HocSinh (chứa thông tin về học sinh):

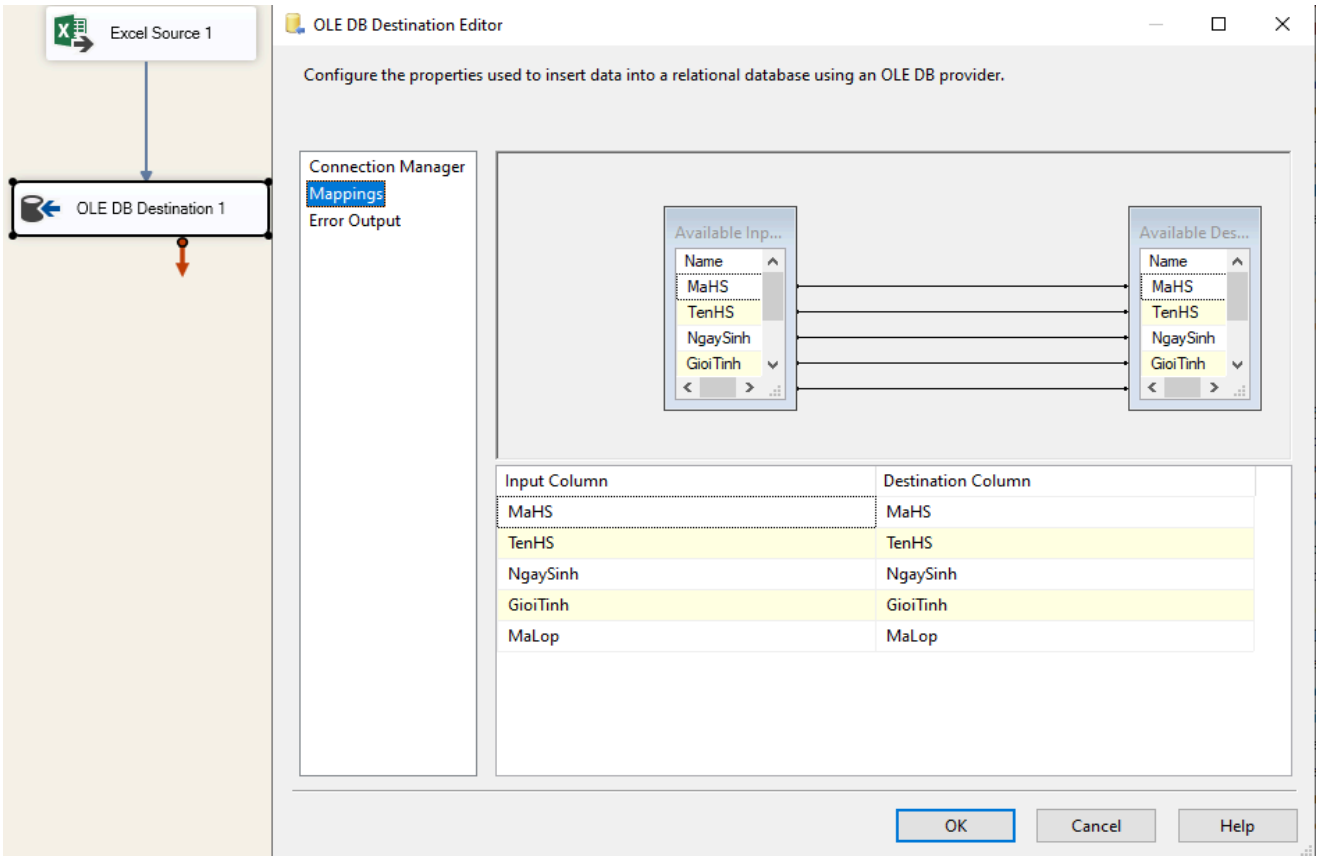
Tên thuộc tính	Ý nghĩa
MaHS	Mã học sinh, là định danh duy nhất cho mỗi học sinh
TenHS	Tên đầy đủ của học sinh
NgaySinh	Ngày sinh của học sinh
GioiTinh	Giới tính của học sinh
MaLop	Mã lớp học mà học sinh thuộc về

2. Extract dữ liệu từ Source vào Stage:

Lấy dữ liệu từ file Excel:

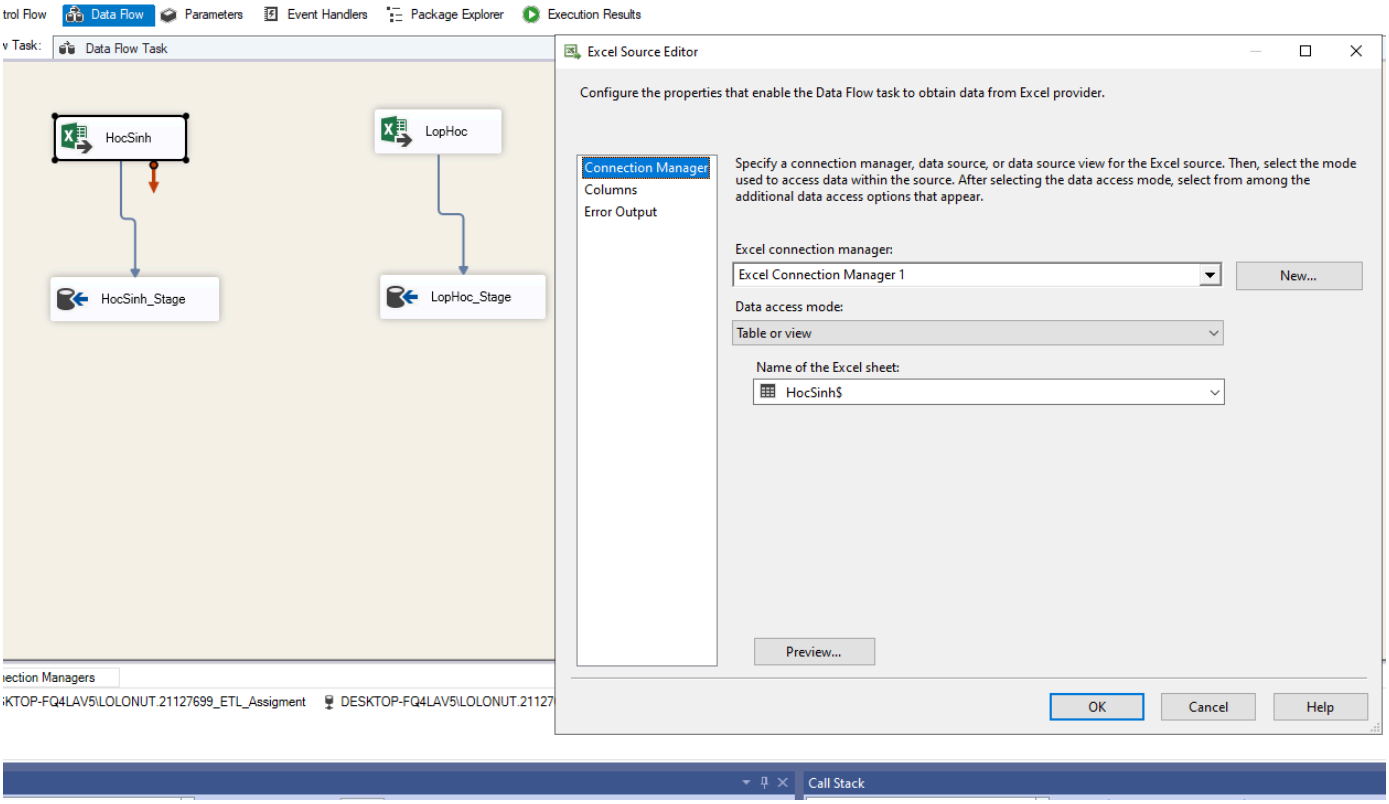
- Sử dụng **Excel Source** để chọn các sheet `Hocsinh` và `LopHoc` làm nguồn dữ liệu.
- **Excel Source**:
  + Kết nối tới file Excel.
  + Chọn Sheet1 (`Hocsinh`). và Sheet2 (`LopHoc`).
  + Cấu hình các cột tương ứng.


Excel Source 1
OLE DB Destination 1
Excel Source Editor
Configure the properties that enable the Data Flow task to obtain data from Excel provider.
Connection Manager
Columns
Error Output
Specify a connection manager, data source, or data source view for the Excel source. Then, select the mode used to access data within the source. After selecting the data access mode, select from among the additional data access options that appear.
Excel connection manager:
Excel Connection Manager
New...
Data access mode:
Table or view
Name of the Excel sheet:
HocSinh$
Preview...
OK
Cancel
Help

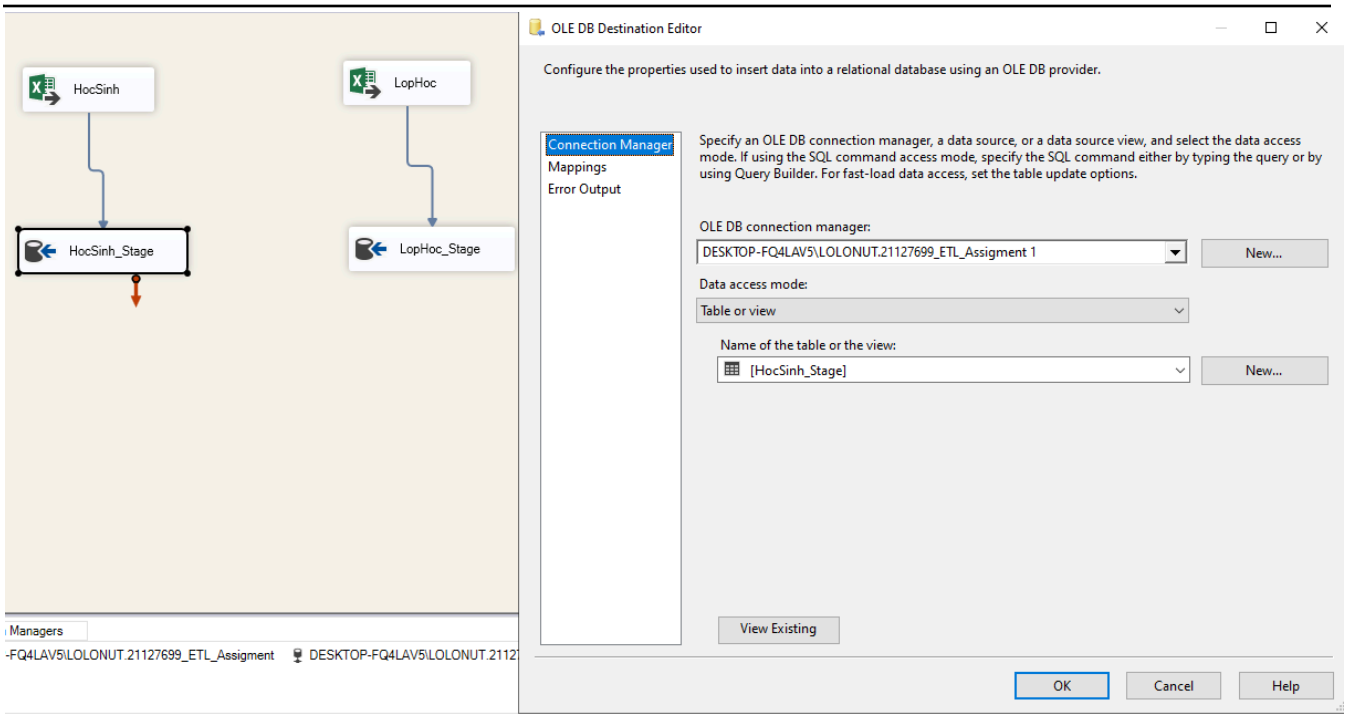

Excel Source 1
OLE DB Destination 1
OLE DB Destination Editor
Configure the properties used to insert data into a relational database using an OLE DB provider.
Connection Manager
Mappings
Error Output
Available Inp...
Name
MaHS
TenHS
NgaySinh
GioiTinh
Available Des...
Name
MaHS
TenHS
NgaySinh
GioiTinh
Input Column
Destination Column
MaHS
MaHS
TenHS
TenHS
NgaySinh
NgaySinh
GioiTinh
GioiTinh
MaLop
MaLop
OK
Cancel
Help

- **Tạo bảng Stage**: Các bảng Stage sẽ có cấu trúc tương tự như nguồn để lưu trữ tạm thời dữ liệu trước khi chuyển vào NDS.

3. Giải thích Flow:



- **Trích xuất dữ liệu từ file Excel:**
  - **HocSinh:** Trích xuất dữ liệu học sinh từ file Excel.
  - **LopHoc:** Trích xuất dữ liệu lớp học từ file Excel.



- Sử dụng **OLE DB Destination** để đưa dữ liệu vào một cơ sở dữ liệu:
  - Chọn một connection manager để kết nối với cơ sở dữ liệu đích. (DESKTOP-FQ4LAV5 LOLONUT.21127699_ETL_Assigment)
  - Chọn "Table or view" để chèn dữ liệu vào một bảng cụ thể trong cơ sở dữ liệu, bảng được đặt tên là "HocSinh_Stage"
- Thực hiện tương tự với LopHoc để tạo bảng "LopHoc_Stage"

II. THIẾT KẾ CẤU TRÚC CÀI ĐẶT:

Lựa chọn kiến trúc cài đặt NDS + DDS, cụ thể như sau:

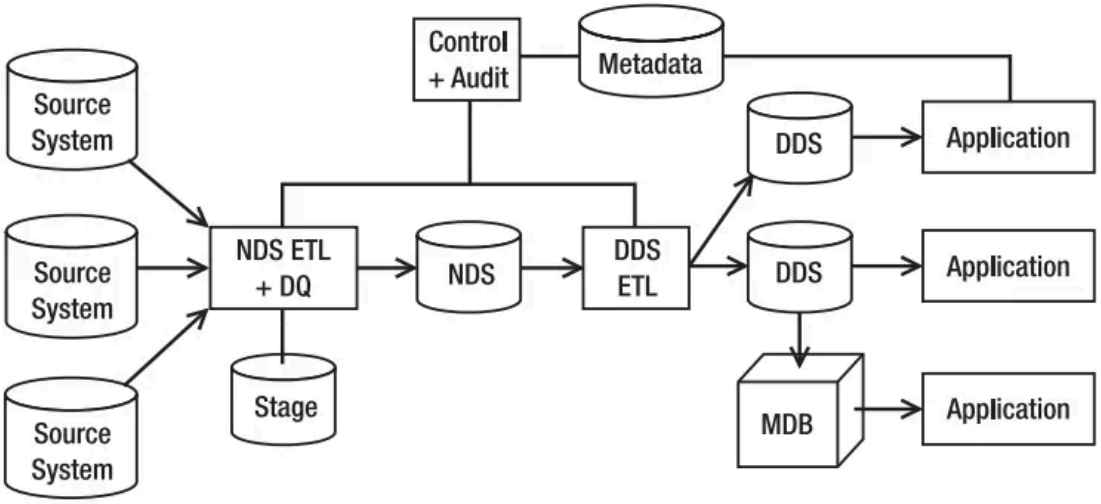


Figure 2-5. *NDS + DDS data flow architecture*

Dữ liệu từ các nguồn sẽ được rút trích và tải vào **Stage** theo phương pháp Incremental trước khi đưa vào **NDS**. Sau khi dữ liệu ở **Stage**, tiến hành **Profiling** để kiểm tra và làm sạch, biến đổi dữ liệu nếu cần. **NDS** là cơ sở dữ liệu chuẩn hóa lưu trữ toàn bộ dữ liệu. Dữ liệu từ **NDS** sẽ được rút trích ra các **DDS** hoặc **Data**

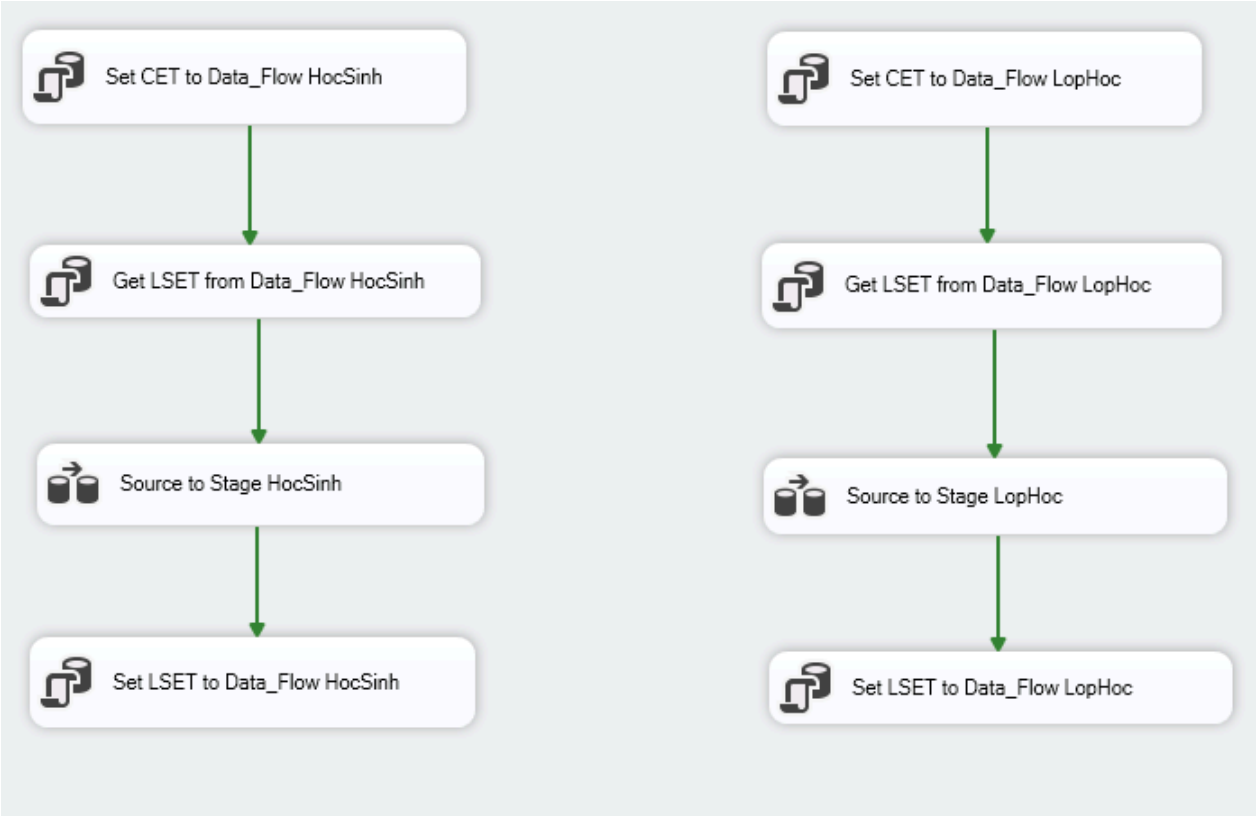
III. GIAI ĐOẠN STAGE:

1. Cấu trúc Data_Flow:

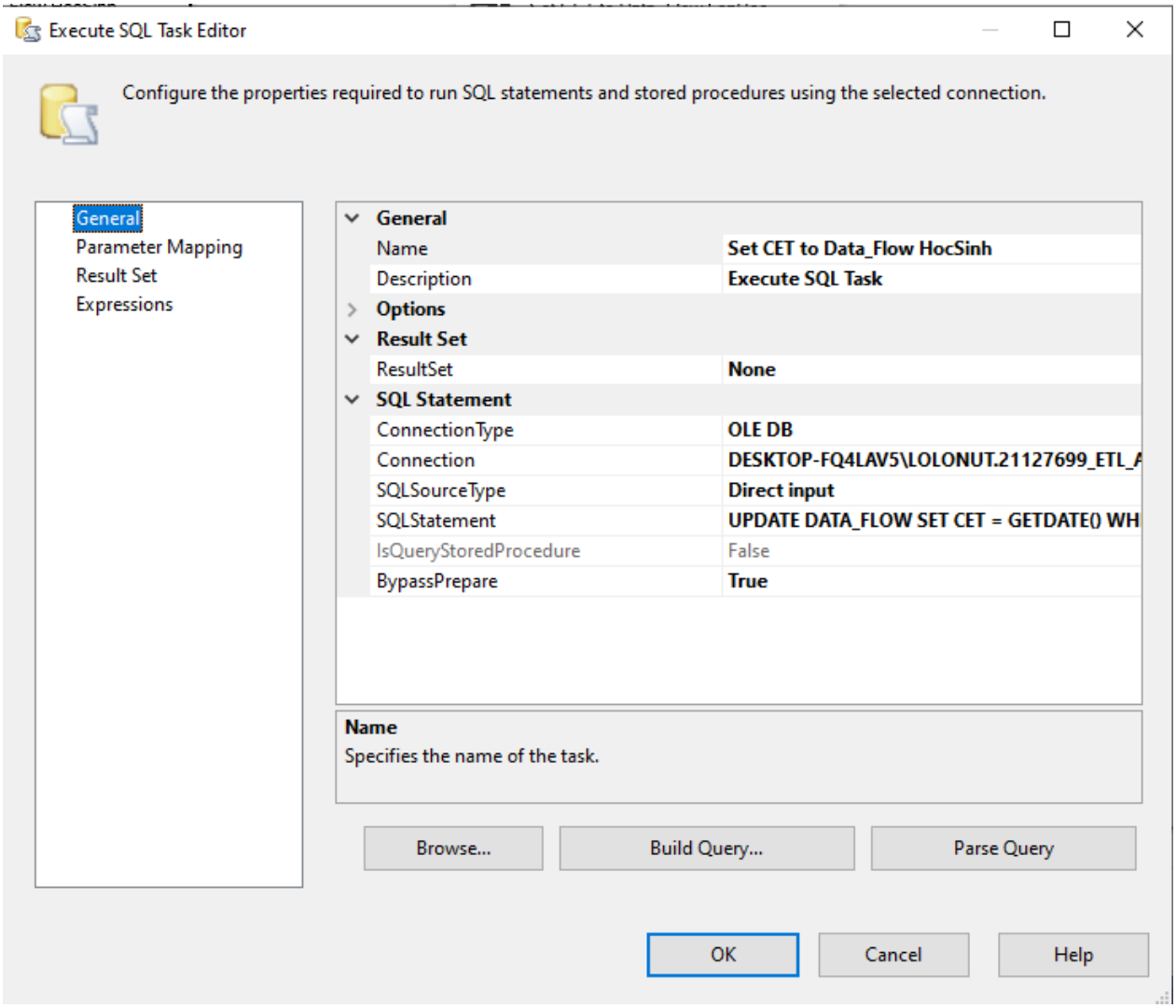
Tên cột	Kiểu dữ liệu	Ý nghĩa	Ghi chú
ID	Số	Khóa chính tự tăng	
Name	Chuỗi	Tên data flow	
LSET	Ngày tháng	Thời gian rút trích thành công gần nhất	
CET	Ngày tháng	Thời gian bắt đầu rút trích	

2. Data Flow:

Để tăng tốc độ rút trích từ nguồn, các bảng ở Stage có cấu trúc tương tự với nguồn. Quá trình ETL sẽ thực hiện như sau:



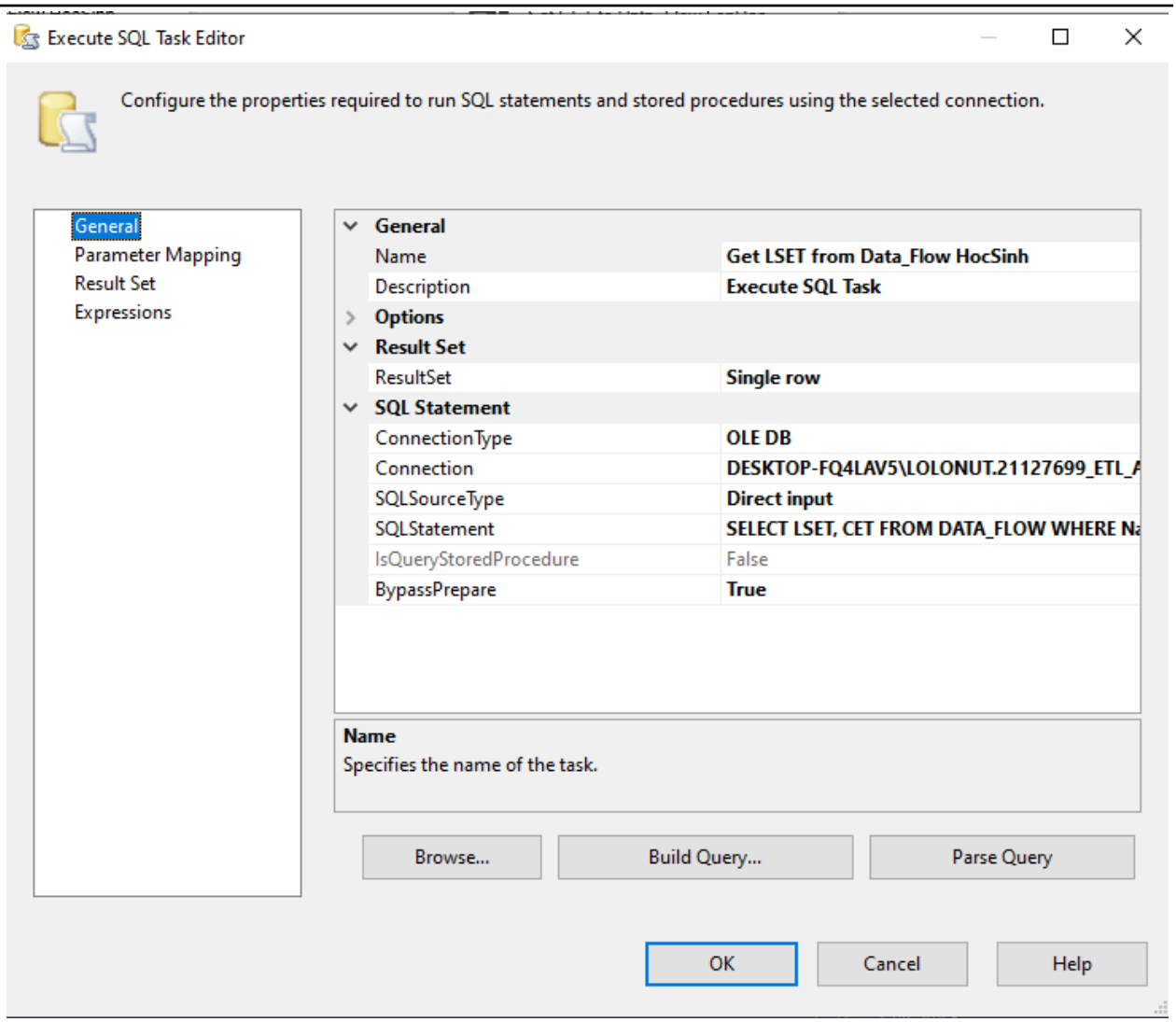
1. Ghi lại **CET** tại Metadata khi bắt đầu ETL từ nguồn vào Stage.



- **Execute SQL Task (Tác vụ Thực thi SQL):** sử dụng để thực thi các câu lệnh SQL trực tiếp vào cơ sở dữ liệu.
- Cập nhật cột "CET" trong bảng "DATA_FLOW" thành giá trị hiện tại (GETDATE()).

```
UPDATE DATA_FLOW SET CET = GETDATE() WHERE Name = 'HocSinh'
```

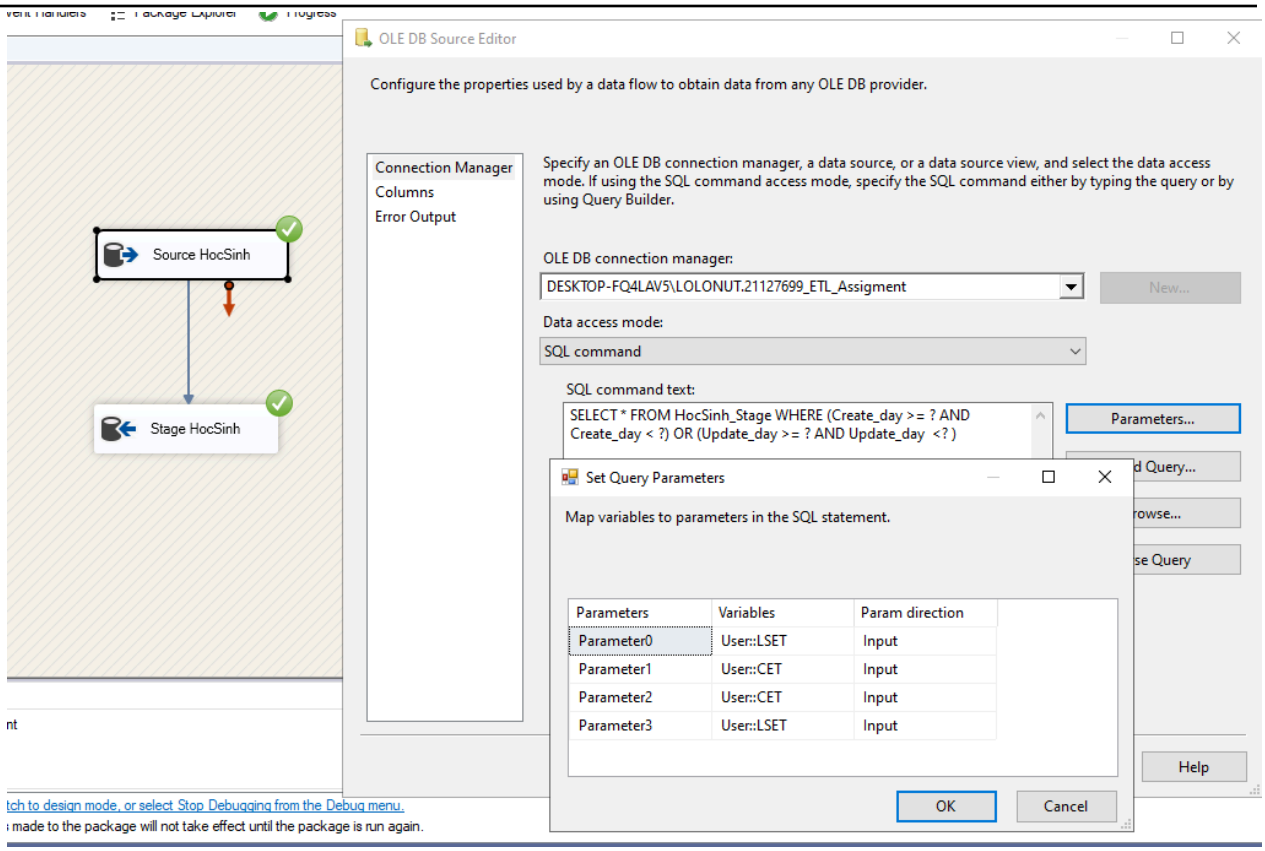
2. Lấy giá trị **LSET** của bảng chuẩn bị nạp vào Stage từ Metadata.



- **Execute SQL Task (Tác vụ Thực thi SQL):** sử dụng để thực thi các câu lệnh SQL trực tiếp vào cơ sở dữ liệu.
- Lấy giá trị của cột "LSET" và "CET" từ bảng "DATA_FLOW"

```
SELECT LSET, CET FROM DATA_FLOW WHERE Name
='HocSinh'
```

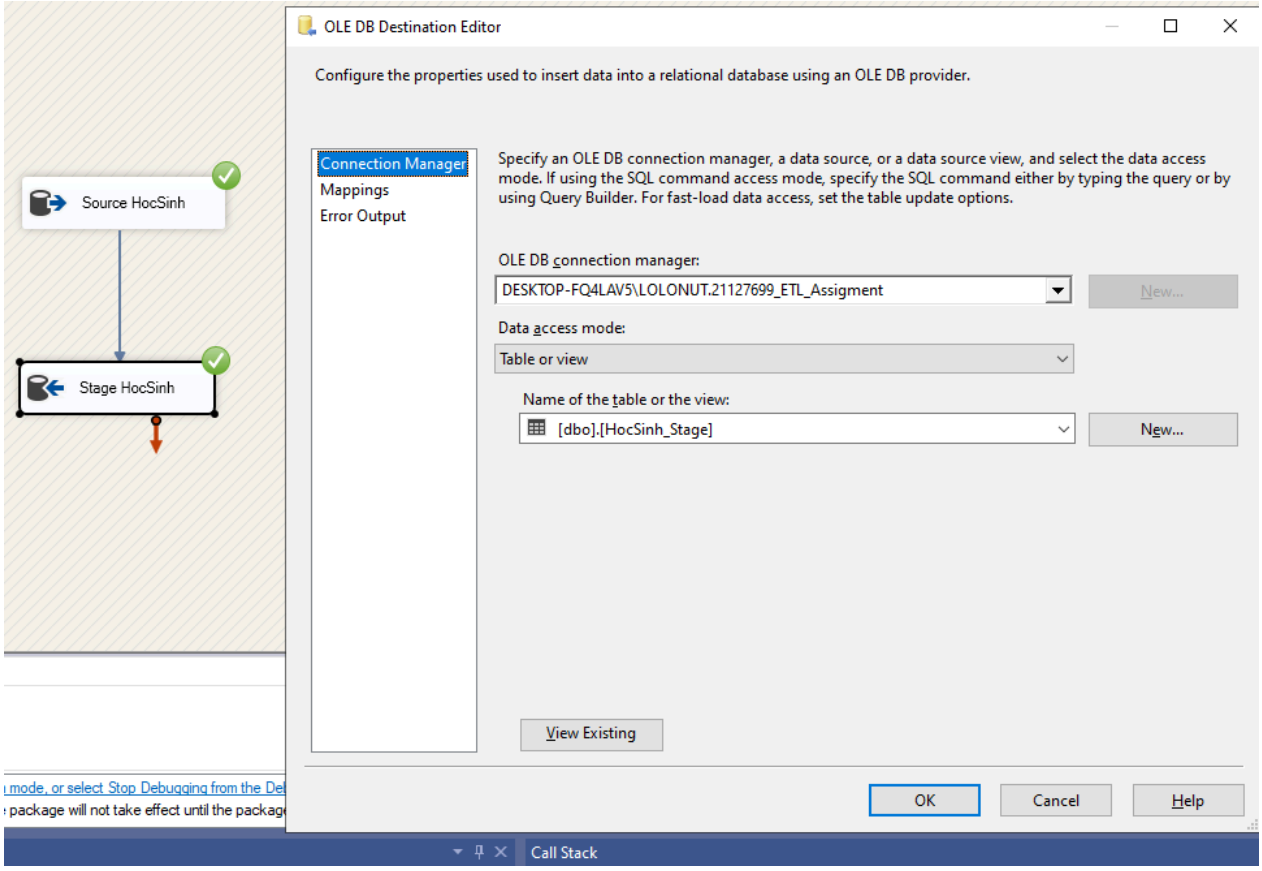
3. Chọn các dòng dữ liệu được tạo ra hoặc cập nhật từ **LSET** đến **CET** từ nguồn và đổ vào Stage.



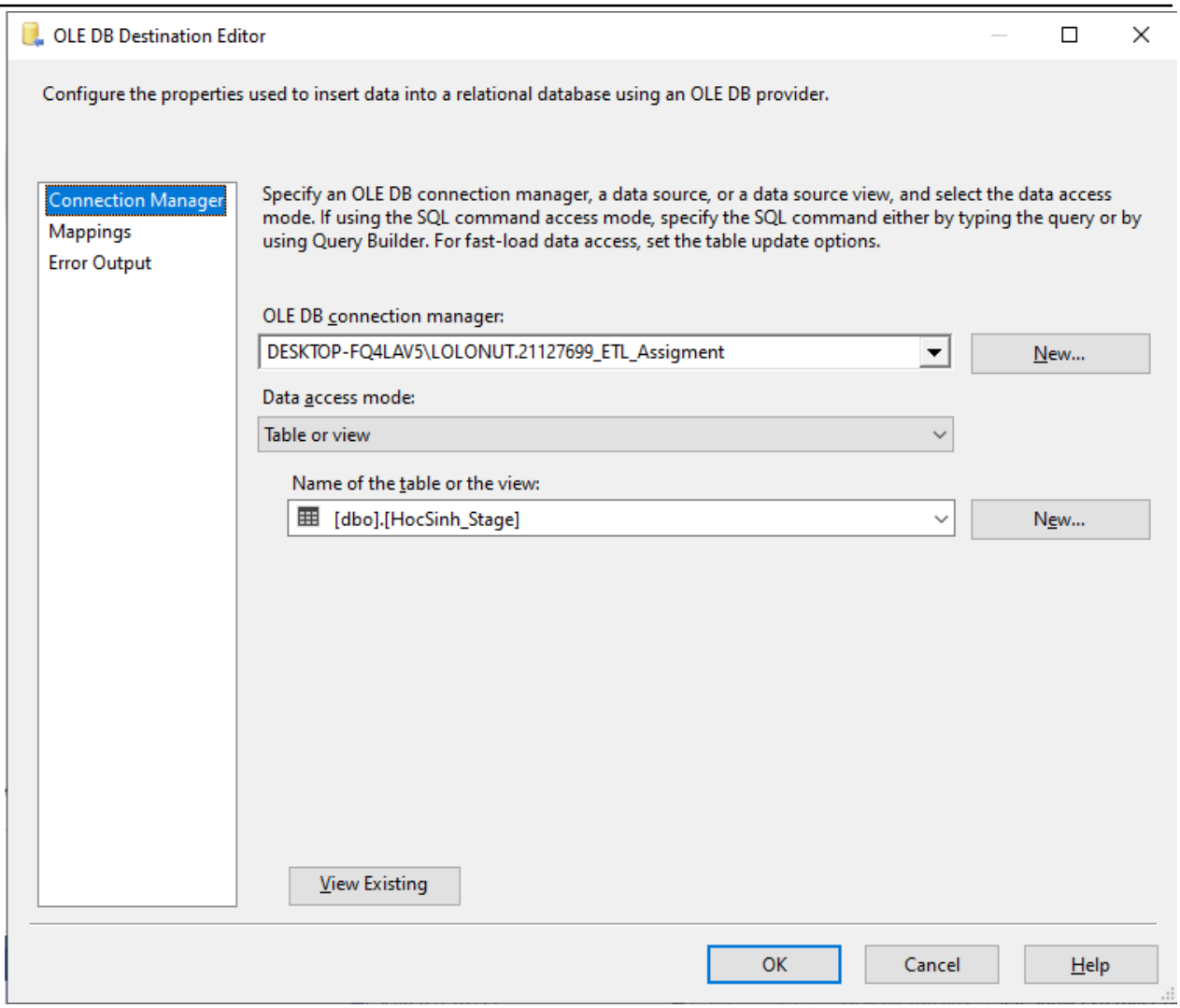
- **OLE DB Source Editor:** để kết nối với cơ sở dữ liệu và trích xuất dữ liệu.
- Chọn tất cả các cột từ bảng "HocSinh_Stage" và chỉ lấy ra các hàng thỏa mãn điều kiện về ngày tạo hoặc ngày cập nhật.

```
SELECT * FROM HocSinh_Stage WHERE (Create_day >= ?
AND Create_day < ?) OR (Update_day >= ? AND
Update_day  <? )
```

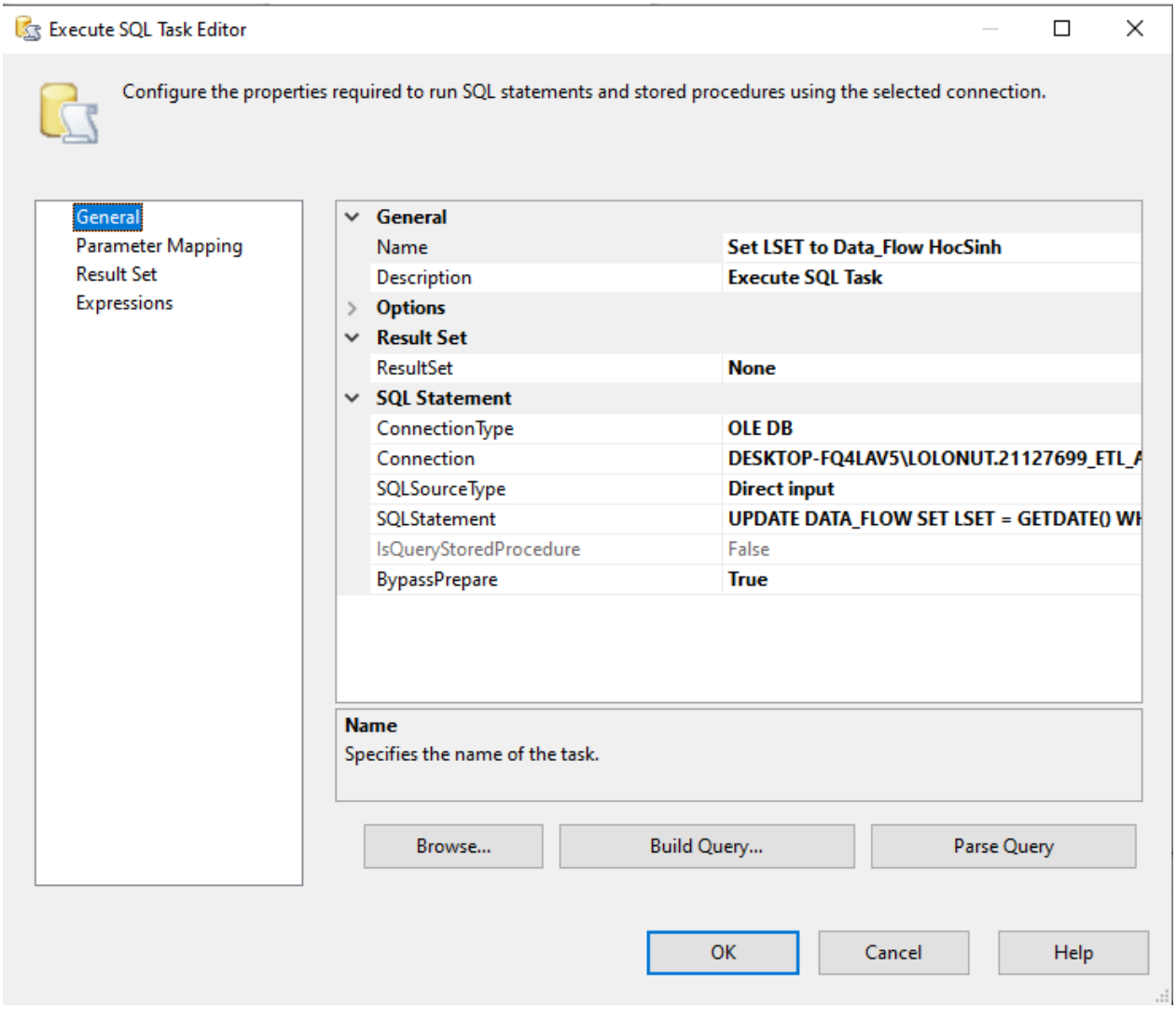
4. Dọn dẹp dữ liệu cũ trong Stage trước khi tải dữ liệu mới.



5. Tiến hành **load** dữ liệu mới vào Stage.



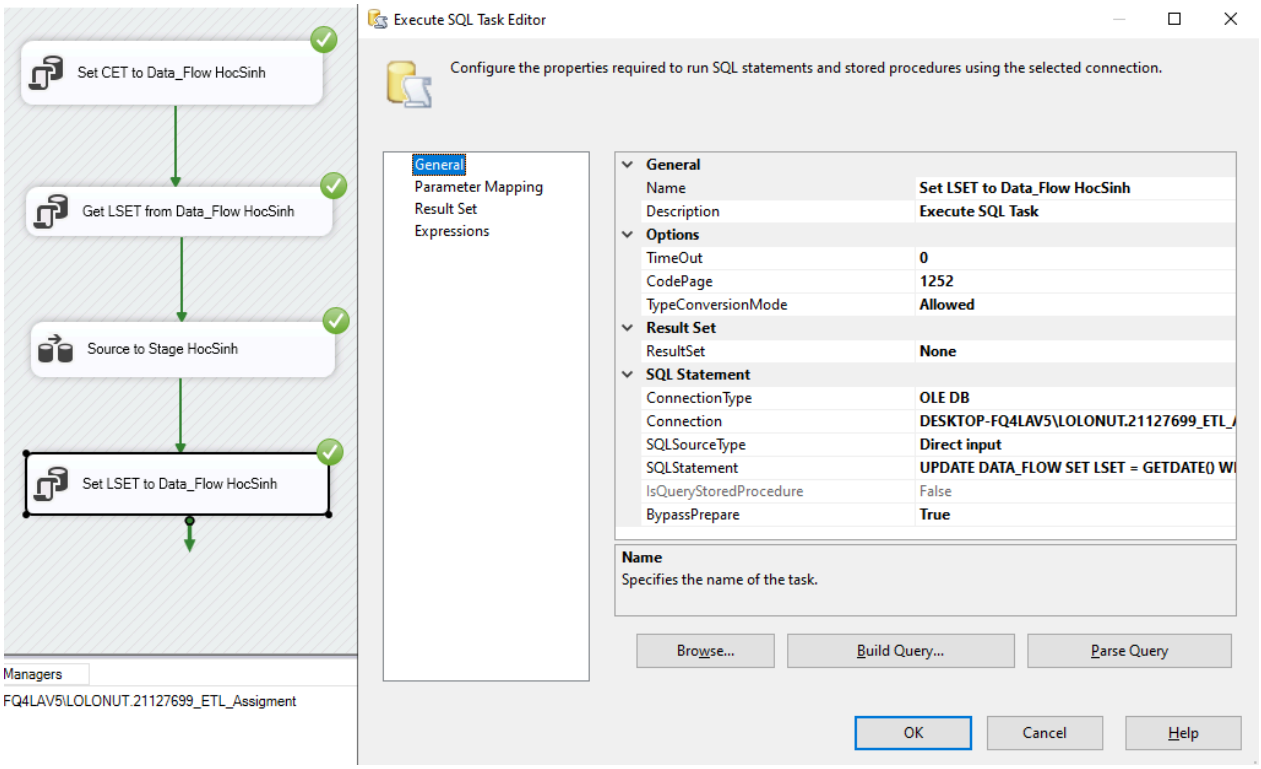
6. **Profiling** dữ liệu vừa được load vào Stage.
7. Cập nhật **LSET** bằng giá trị CET hiện tại vào Metadata.



- **Execute SQL Task (Tác vụ Thực thi SQL):** sử dụng để thực thi các câu lệnh SQL trực tiếp vào cơ sở dữ liệu.
- Cập nhật "LSET" trong bảng "DATA_FLOW" thành giá trị hiện tại (GETDATE()).

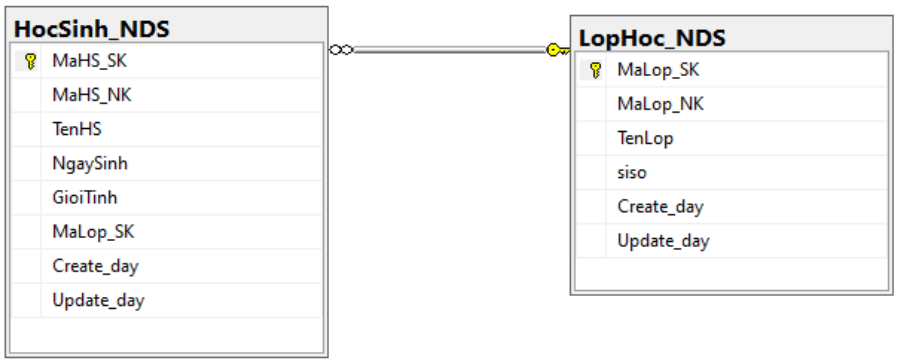
```
UPDATE DATA_FLOW SET LSET = GETDATE() WHERE Name
='HocSinh'
```

Thực hiện tương tự với LopHoc để đổ dữ liệu vào bảng "LopHoc_Stage"



IV. GIAI ĐOẠN : STAGE → NDS

1. Cấu trúc NDS:



2. Chi tiết từng bảng trong NDS:

a. Bảng LopHoc_NDS:

STT	Tên thuộc tính	Mô tả	Transformation rules	Nguồn	SCD
1	MaLop_SK	Khóa tự tăng	Tạo khóa mới		1
2	MaLop_NK	Khóa tự nhiên		Nguồn: Stage_LopHoc	1
3	TenLop	Tên lớp học		Nguồn: Stage_LopHoc	1
4	Siso	Sĩ số (số lượng học sinh)	Trường được tính toán	Tính từ HocSinh_NDS	1
5	Create_day	Thời gian tạo bản ghi	Thiết lập GETDATE()		1
6	Update_day	Thời gian cập nhật bản ghi	Thiết lập GETDATE()		1

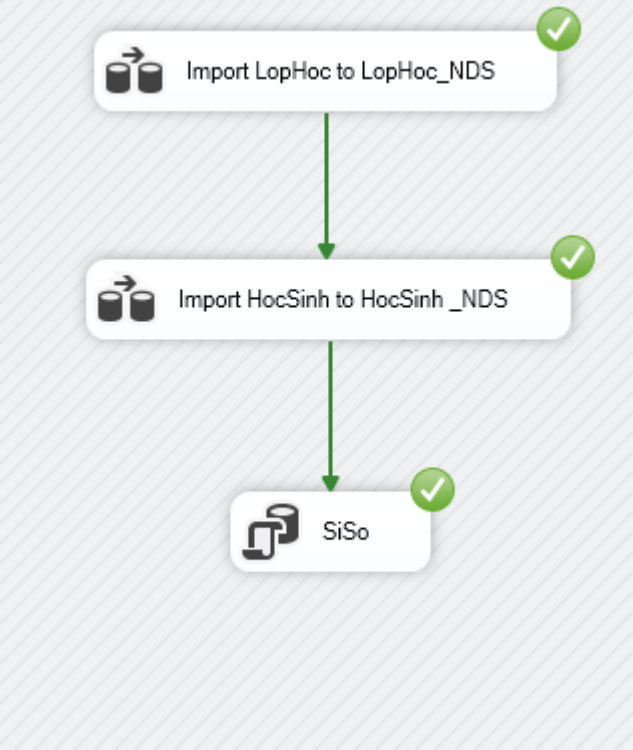
b. Bảng HocSinh_NDS:

STT	Tên thuộc tính	Mô tả	Transformation rules	Nguồn	SCD
1	MaHS_SK	Khóa tự tăng	Tạo khóa mới		1
2	MaHS_NK	Khóa tự nhiên		Nguồn: Stage_HocSinh	1
3	MaHS_NK	Tên của học sinh		Nguồn: Stage_HocSinh	1
4	NgaySinh	Ngày sinh của học sinh	Trường được tính toán	Nguồn: Stage_HocSinh	1
5	GioiTinh	Giới tính của học sinh		Nguồn: Stage_HocSinh	1
6	MaLop_SK	Khóa ngoại tham chiếu tới	Tra cứu và ánh xạ từ nguồn	Nguồn: Stage_HocSinh	1

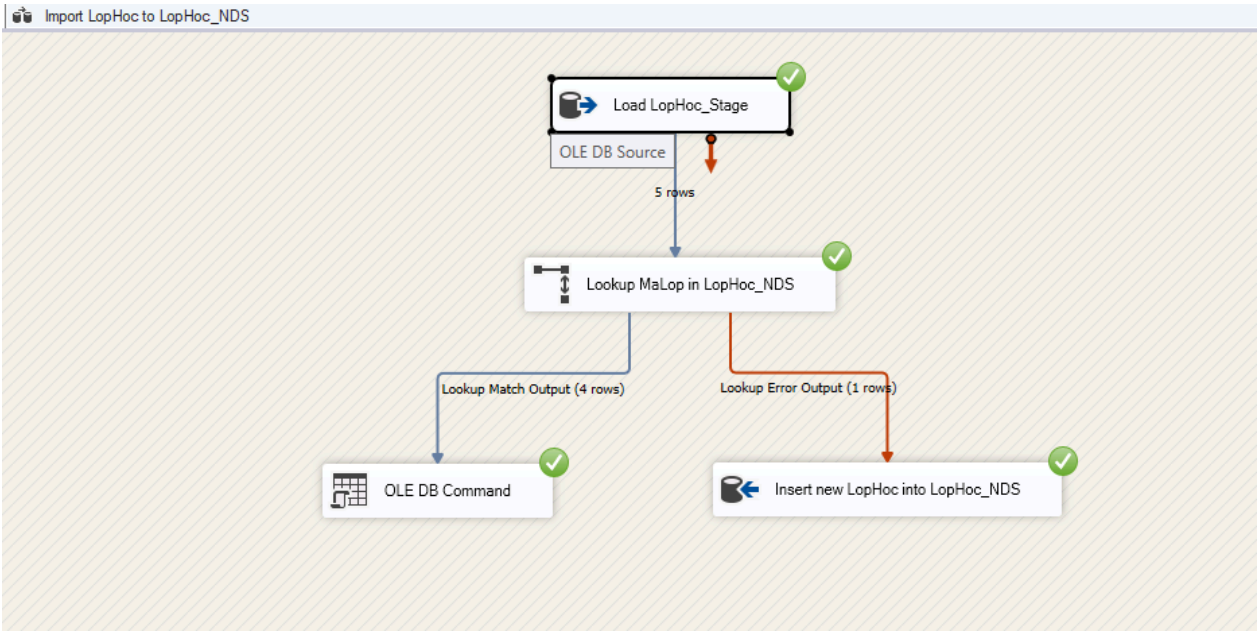
		bảng LopHoc_NDS			
7	Create_day	Thời gian tạo bản ghi	Thiết lập GETDATE()		1
8	Update_day	Thời gian cập nhật bản ghi	Thiết lập GETDATE()		1

3. Giải thích Flow:

a. Control Flow:

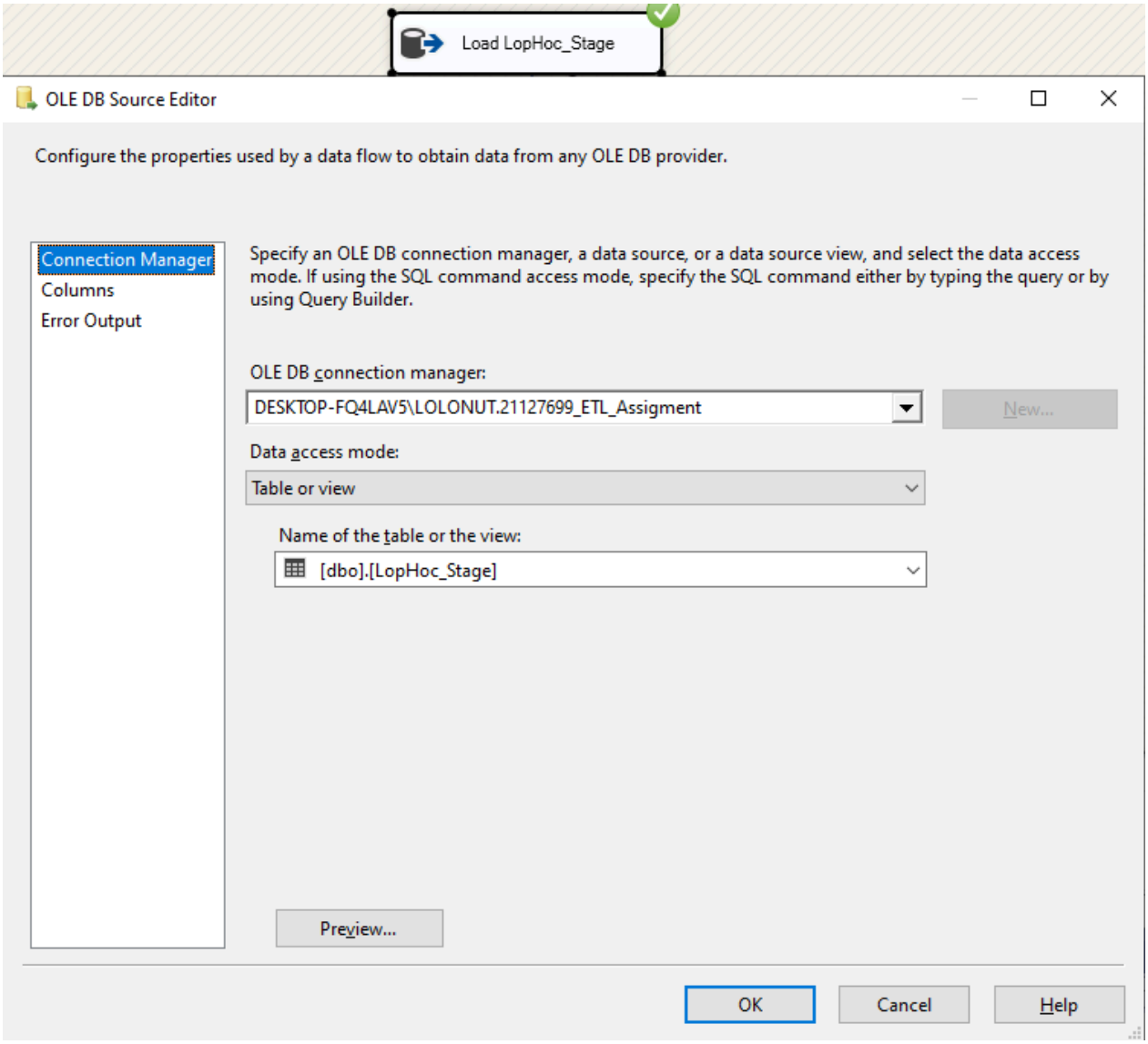


b. Data Flow:

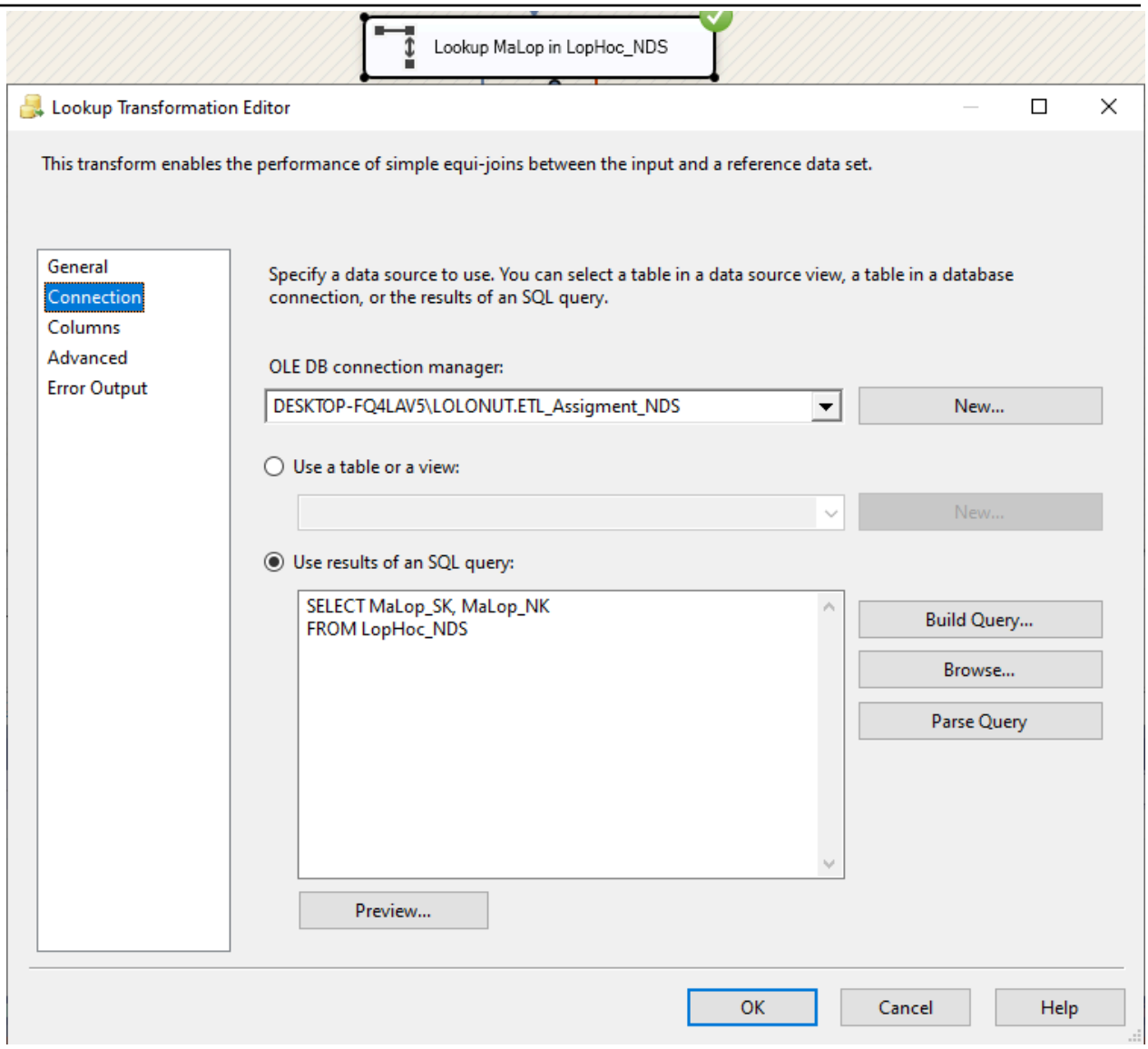


1. ***<u>Import LopHoc to LopHoc_NDS</u>***

- **"OLE DB Source":** Xuất dữ liệu từ một cơ sở dữ liệu và đưa vào SSIS.



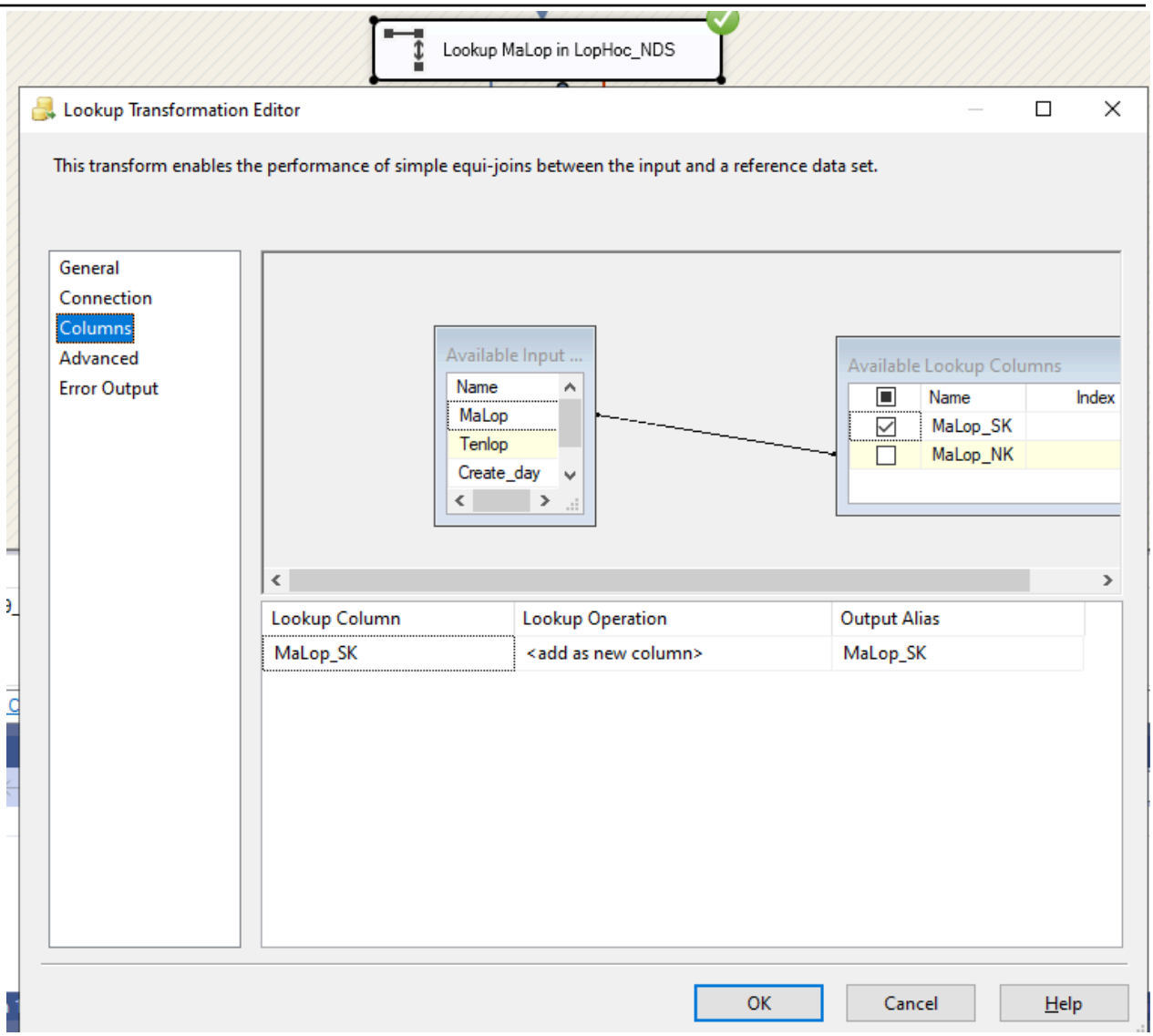
- Sử dụng **OLE DB Destination** để đưa dữ liệu vào một cơ sở dữ liệu:
  - Chọn một connection manager. (DESKTOP-FQ4LAV5 LOLONUT.21127699_ETL_Assigment) - Chứa bảng Stage
  - Chọn "LopHoc_Stage" để trích xuất dữ liệu ra xử lý.



- **"Lookup Transformation".**: Tham chiếu khóa SK - NK
  - Chọn nguồn dữ liệu tham chiếu.

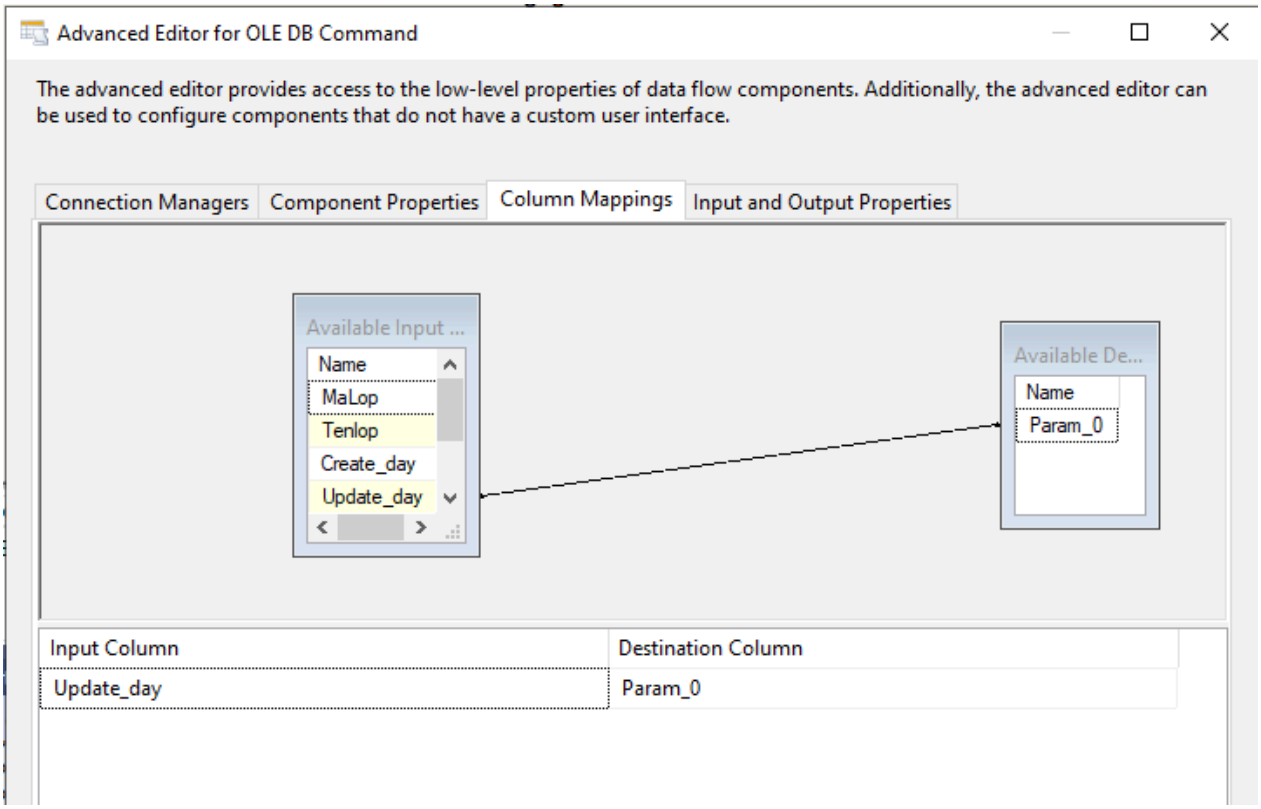
```
SELECT MaLop_SK, MaLop_NK
FROM LopHoc_NDS
```

  - Mapping dữ liệu SK-NK

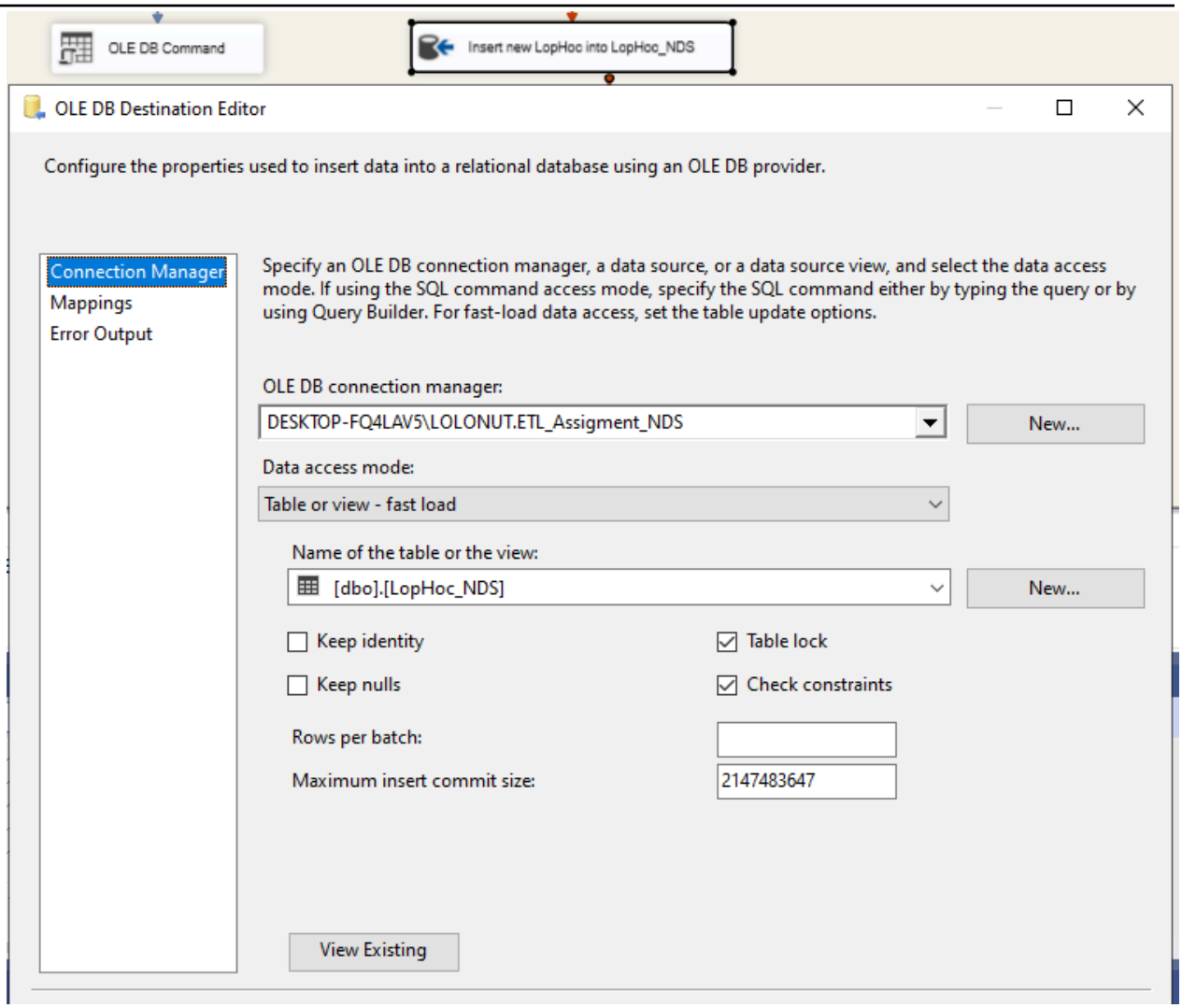


- **Advanced Editor for OLE DB Command** tùy chỉnh sâu các thuộc tính của một thành phần Data Flow cụ thể, trong trường hợp này là OLE DB Command.
  - Cập nhật trường Update_day trong bảng LopHoc_NDS

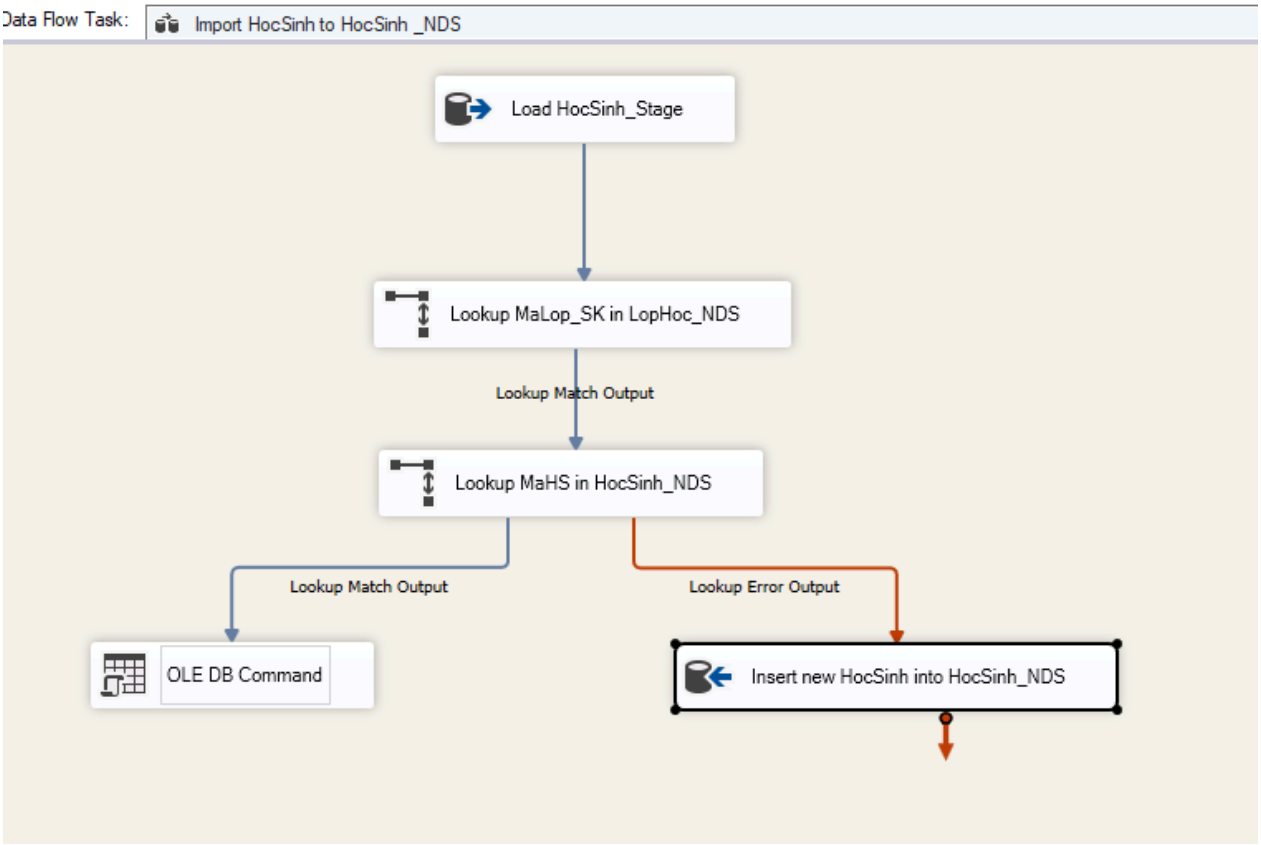
```
UPDATE LopHoc_NDS set Update_day =?
```



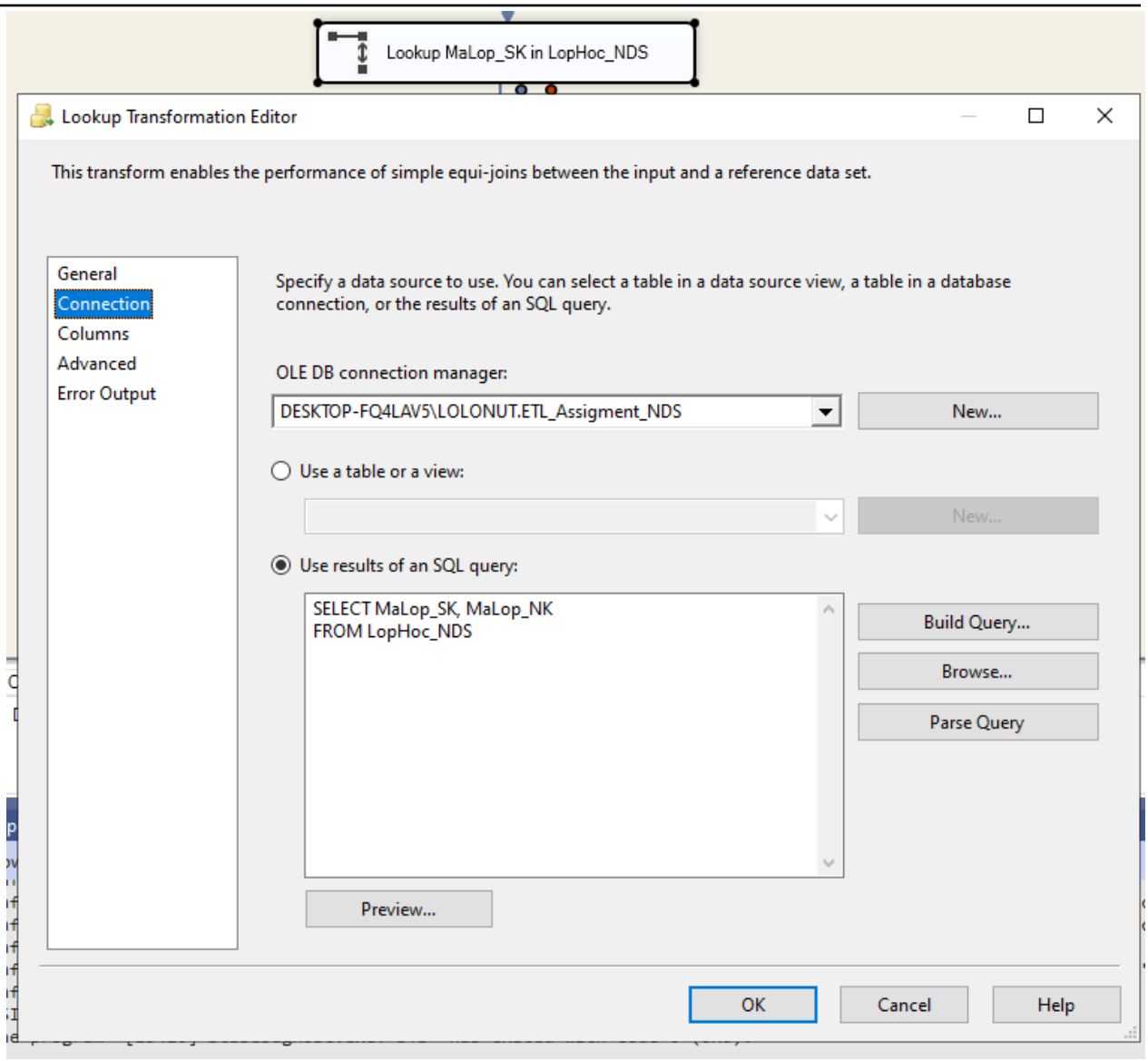
- Sử dụng **OLE DB Destination** để đưa dữ liệu vào một cơ sở dữ liệu: Ở đây nếu như dữ liệu không được tìm thấy (Theo luồng ở trên) thì sẽ vào bảng NDS để thêm trường mới.
  - **[dbo].[LopHoc_NDS]**: bảng trong cơ sở dữ liệu muốn chèn dữ liệu vào.


OLE DB Command
Insert new LopHoc into LopHoc_NDS
OLE DB Destination Editor
Configure the properties used to insert data into a relational database using an OLE DB provider.
Connection Manager
Mappings
Error Output
Specify an OLE DB connection manager, a data source, or a data source view, and select the data access mode. If using the SQL command access mode, specify the SQL command either by typing the query or by using Query Builder. For fast-load data access, set the table update options.
OLE DB connection manager:
DESKTOP-FQ4LAV5\LOLONUT.ETL_Assigment_NDS
New...
Data access mode:
Table or view - fast load
Name of the table or the view:
[dbo].[LopHoc_NDS]
New...
Keep identity
Table lock
Keep nulls
Check constraints
Rows per batch:
Maximum insert commit size:
2147483647
View Existing

2. ***Import HocSinh to HocSinh _NDS***



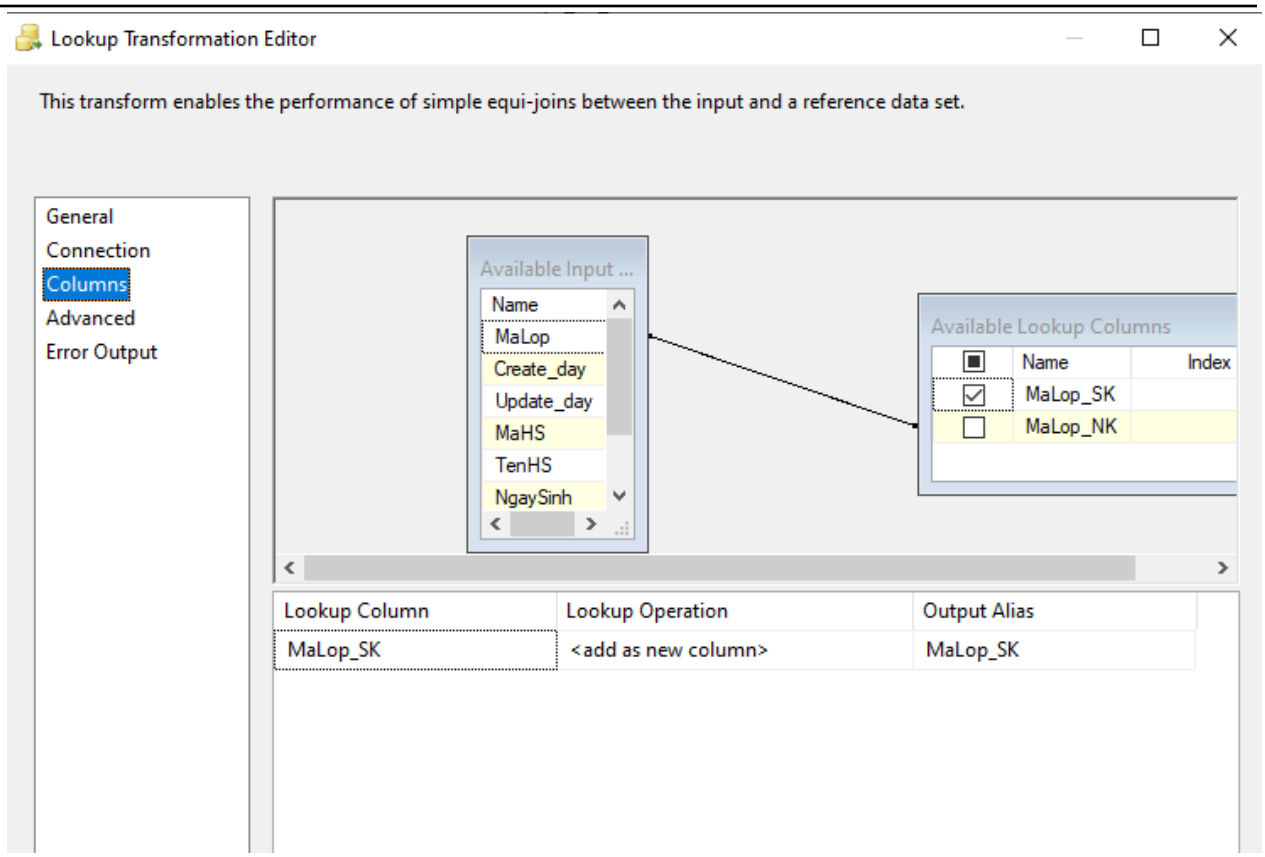
Thực hiện tương tự ***Import LopHoc to LopHoc_NDS*** , tuy nhiên do trong bảng HocSinh_NDS còn bao gồm 1 khóa ngoại là "MaLop_SK", nên chúng ta cần phải thực hiện thêm 1 lần LookUp **"Lookup MaLop_SK in LopHoc_NDS"**



- **"Lookup Transformation".**: Tham chiếu khóa SK - NK
  - Chọn nguồn dữ liệu tham chiếu.

```
SELECT MaLop_SK, MaLop_NK
FROM LopHoc_NDS
```

  - Mapping dữ liệu SK-NK


Lookup Transformation Editor
This transform enables the performance of simple equi-joins between the input and a reference data set.
General
Connection
Columns
Advanced
Error Output
Available Input ...
Name
MaLop
Create_day
Update_day
MaHS
TenHS
NgaySinh
Available Lookup Columns
Name
Index
MaLop_SK
MaLop_NK
Lookup Column
Lookup Operation
Output Alias
MaLop_SK
<add as new column>
MaLop_SK

TÀI LIỆU THAM KHẢO

Công cụ và phần mềm hỗ trợ:

STT	Chức năng	Công cụ
[1]	Báo cáo	Google Docs
[2]	Quay Video demo	OBS Open Broadcaster Software
[3]	Thiết kế SSIS	Microsoft Visual Studio 2022
[4]	Quản lý dự án	SSMS 20
[5]	Thiết kế CSDL	SQL sever

Tài liệu tham khảo:

[1]Tài liệu môn học Hệ thống thông tin phục vụ trí tuệ kinh doanh - 21HTTT2

[2] https://github.com/buicongdanh/BI_DATH/blob/main/Script/nds.sql

[3] https://www.youtube.com/watch?v=TXqDN1aIdjo&t=376s

[4] https://www.youtube.com/watch?v=UTMpdVfNYV4&t=229s