Mục lục

I. Tổng quan dự án

Ở dự án này, chúng ta sẽ xây dựng một Data Warehouse từ một tập dữ liệu bao gồm thông tin quản lý thú cưng trên toàn nước Mỹ (PetFinder.csv). Data Warehouse sẽ cần đảm bảo các yêu cầu sau:

- Thiết kế được ERD cho Data Warehouse dựa trên bộ dữ liệu cho trước.
- Xác định được tối thiểu 3 business queries (truy vấn nghiệp vụ) để thực hiện phân tích dữ liệu.
- Xây dựng được quy trình ETL trong SSIS.
- Mô tả được quy trình ETL để đưa dữ liệu vào Data Warehouse.
- Đưa được dữ liệu vào Database.
- Viết các câu lệnh SQL để xây dựng Database.
- Viết các câu lệnh SQL để lấy dữ liệu cho các business queries đã xác định.

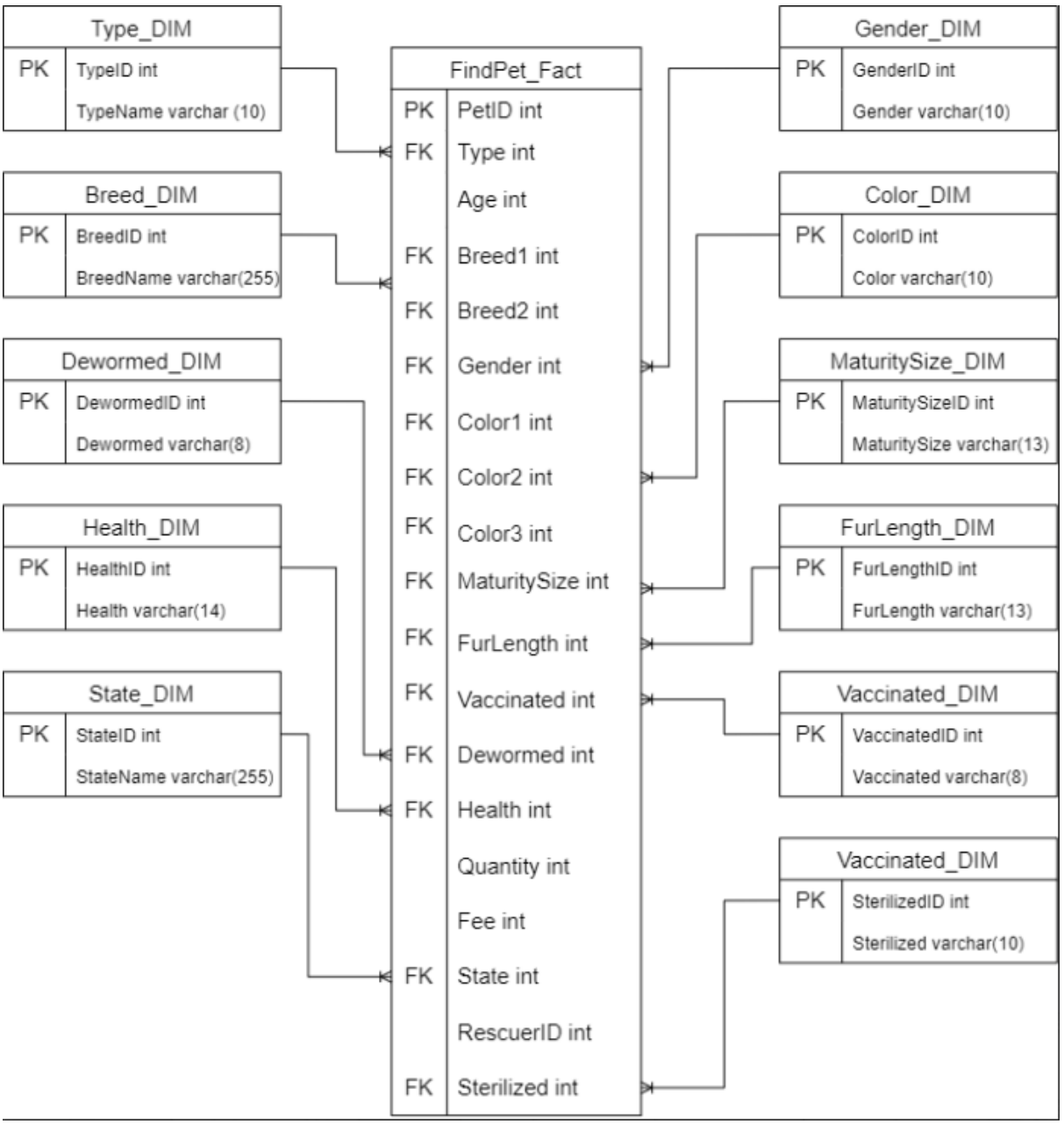
II. Yêu cầu và thực hiện dự án

A. Thiết kế ERD

ERD bao gồm 1 Fact table giống với bảng dữ liệu trong file csv nguồn (FindPet.csv) và 11 Dim table cho các trường:

1. Type
2. Breed
3. Gender
4. Color
5. MaturitySize
6. FurLength
7. Vaccinated
8. Dewormed
9. Sterilized
10. Health
11. State

Như hình dưới đây:



Lược đồ ERD

B. Tạo database (*file SQL Create Database)

Đầu tiên ta sẽ tạo các Dim table được liệt kê ở trên theo đúng định dạng trong lược đồ ER. Sau đó ta tạo 1 Fact table chứa các trường giống với file csv nguồn và tạo 1 staging layer table để staging dữ liệu trước khi đưa dữ liệu vào Fact table.

C. Các truy vấn nghiệp vụ

1. Có bao nhiêu con vật nuôi trên 100 tuổi

```sql
-- 1.Có bao nhiêu con vật nuôi trên 100 tuổi
select * from FindPet_Fact
where Age > 100
```

90 %

Results | Messages

	PetID	Type	Age	Breed1	Breed2	Gender	Color1	Color2	Color3	MaturitySize	Sterilized	FurLength	Vaccinated	Dewormed	Health	Quantity	Fee	State	RescuerID
1	113	54	120	115	120	50	1	2	7	2	2	4	2	2	1	1	0	4	2280777
2	145	54	120	23	146	50	2	6	6	3	1	4	3	3	1	2	0	13	4470292
3	513	1	132	66	120	1	1	2	6	3	3	4	3	3	1	1	0	13	4344280
4	1057	54	120	160	120	1	4	5	6	5	2	2	2	2	1	1	0	13	3530935
5	1141	54	120	115	120	1	1	7	6	3	2	2	2	2	2	1	0	4	3766388
6	1363	54	144	185	120	50	3	6	6	3	1	2	2	2	1	1	0	13	9850013
7	1703	1	120	99	120	1	2	3	7	3	3	2	2	2	1	1	0	13	8143715
8	1858	54	132	115	114	1	1	6	6	5	1	4	3	3	1	1	0	13	7377962
9	1867	54	135	146	120	50	5	7	6	5	2	4	2	3	2	1	0	13	4135621
10	2263	1	120	66	120	1	1	6	6	5	3	4	2	2	1	1	0	4	6334475
11	2497	54	132	97	120	1	2	7	6	5	1	4	3	3	1	1	0	13	6083869
12	2767	54	108	23	120	50	1	2	7	3	1	2	3	1	1	1	0	4	5008341
13	2830	54	120	129	120	1	5	7	6	5	3	2	2	2	1	1	0	13	327766
14	3195	54	120	115	115	1	2	6	6	3	3	2	1	1	1	1	0	6	5387680
15	3684	54	108	86	120	50	4	6	6	2	2	2	3	3	1	1	0	4	1035449
16	3931	1	132	66	120	50	3	7	6	3	3	4	3	3	1	1	0	13	1584716
17	3998	1	212	66	120	1	1	6	6	3	1	4	3	3	1	1	0	9	5997594
18	4071	54	120	142	120	50	1	6	6	2	1	1	3	3	1	1	0	4	5642497
19	4344	54	108	135	120	1	7	6	6	5	3	1	3	3	1	1	200	13	4088065
20	4456	54	120	114	120	1	1	6	6	5	1	4	3	3	2	1	0	13	5321553
21	4522	54	108	23	120	1	4	3	6	3	1	4	3	3	1	1	0	10	5773692
22	4617	54	108	54	54	50	2	6	6	5	3	1	3	3	1	1	0	9	469698

2. Ở bang nào có số lượng vật nuôi lớn nhất

```sql
-- 2.Ở bang nào có số lượng vật nuôi lớn nhất
select
    top 1 [SL_vatnuoi] = count (PetID) ,
    b.StateName
from FindPet_Fact a
join State_DIM b on b.StateID = a.State
    group by b.StateName
    order by count (PetID) Desc

-- 3. Tỉ lệ vật nuôi được tiêm phòng là bao nhiêu
select
    [Vaccinated] = count (a.Vaccinated),
```

90 %

Results | Messages

	SL_vatnuoi	StateName
1	8712	Selangor

3. Tỉ lệ vật nuôi được tiêm phòng là bao nhiêu

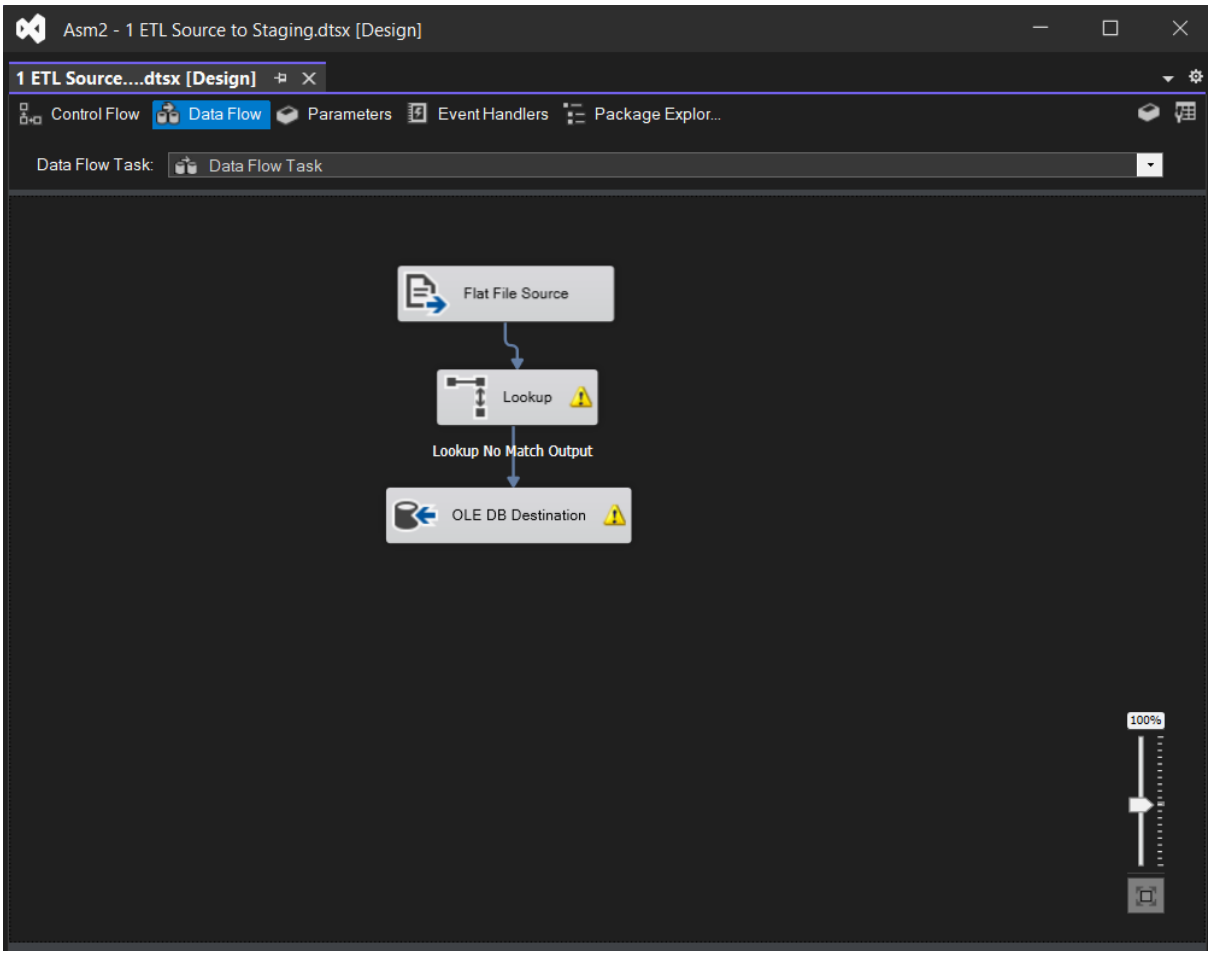
```sql
-- 3. Tỉ lệ vật nuôi được tiêm phòng là bao nhiêu
select
    [Vaccinated] = count (a.Vaccinated),
    [TotalPet] = count (a.TotalPet),
    [VaccinatedRate] = cast ( ((cast( count (a.Vaccinated) as float) / cast( count (a.TotalPet) as float) ) as decimal (10,2) ) -- Tính tỉ lệ tiêm chủng
from
(
    select
        [Vaccinated] = case when Vaccinated = 3 then COUNT(PetID) over (partition by PetID) else null end, -- Tính số vật nuôi được tiêm phòng
        [TotalPet] = count(PetID) over (partition by PetID) -- Tính tổng số vật nuôi
    from FindPet_Fact
) a

update FindPet_Fact
```

90 %

Results | Messages

	Vaccinated	TotalPet	VaccinatedRate
1	5899	14993	0.39

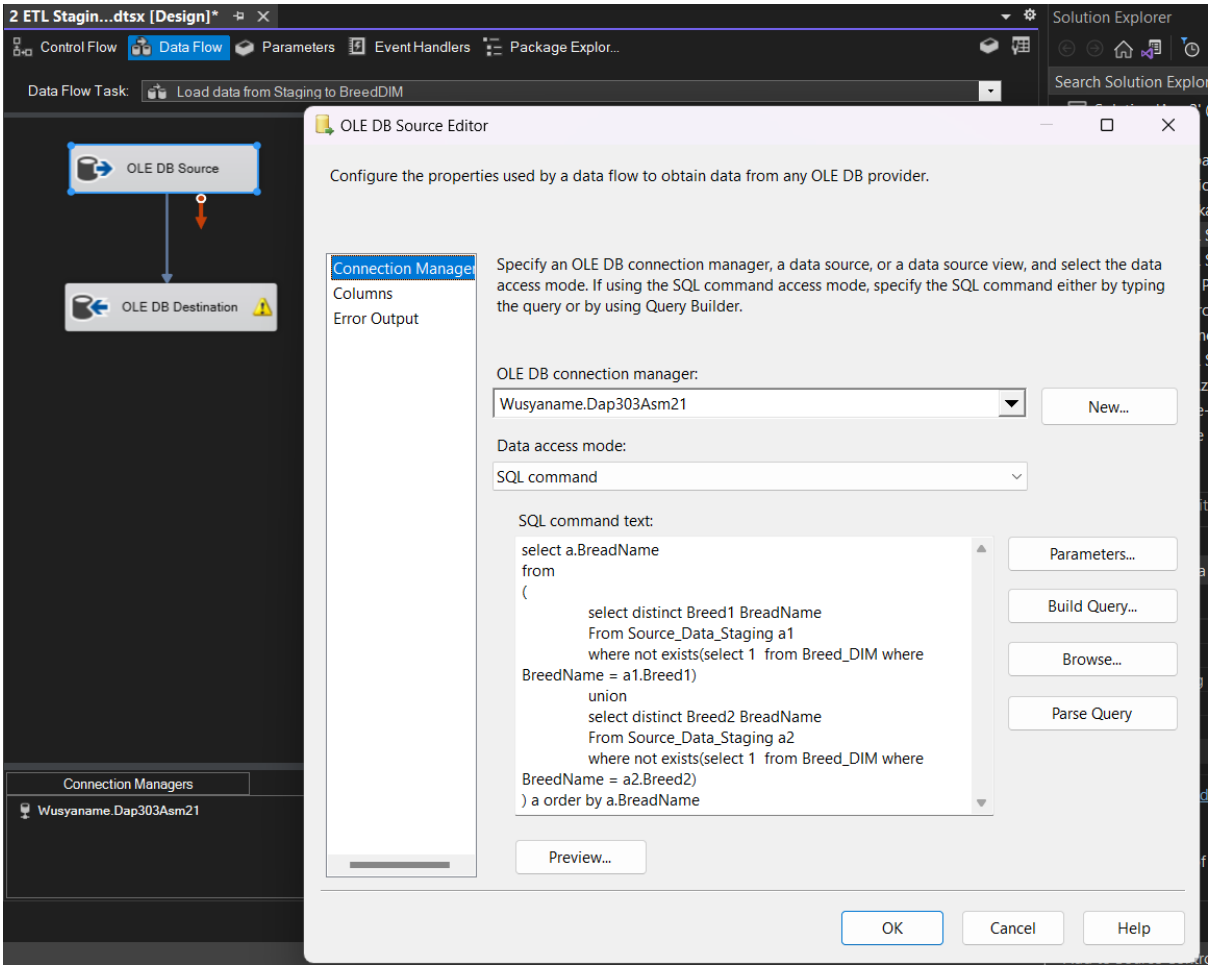
D. Xây dựng ETL cho Staging table



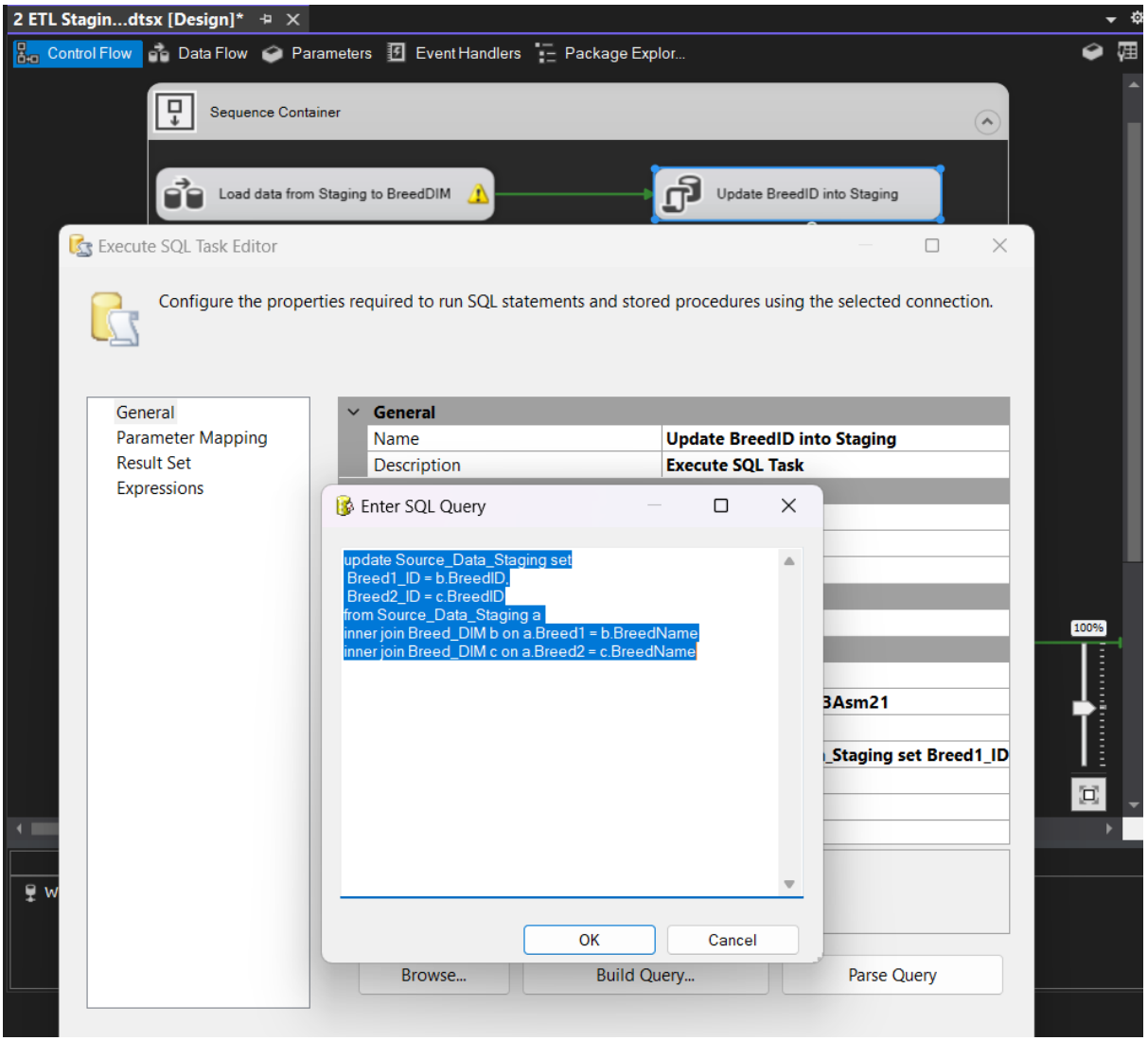
Như hình ta có thể thấy chúng ta sẽ tạo một control flow để đưa dữ liệu từ nguồn csv(PetFinder) vào staging table (Source_Data_Staging) để chuẩn bị dữ liệu trước khi đưa dữ liệu vào Fact.
Sử dụng Lookup để không chèn dữ liệu đã có trong staging.

E. Xây dựng ETL cho Dimension table

Đối với từng Dim table chúng ta sẽ lấy dữ liệu từ nguồn là Staging table để truyền dữ liệu vào từng Dim table tương ứng. Ví dụ sau đây là ETL cho Breed_Dim


2 ETL Stagin...dtsx [Design]*
Control Flow
Data Flow
Parameters
Event Handlers
Package Explor...
Data Flow Task:
Load data from Staging to BreedDIM
OLE DB Source
OLE DB Destination
Connection Managers
Wusyaname.Dap303Asm21
Solution Explorer
Search Solution Explor
OLE DB Source Editor
Configure the properties used by a data flow to obtain data from any OLE DB provider.
Connection Manager
Columns
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.
OLE DB connection manager:
Wusyaname.Dap303Asm21
New...
Data access mode:
SQL command
SQL command text:
select a.BreadName
from
(
select distinct Breed1 BreadName
From Source_Data_Staging a1
where not exists(select 1 from Breed_DIM where
BreedName = a1.Breed1)
union
select distinct Breed2 BreadName
From Source_Data_Staging a2
where not exists(select 1 from Breed_DIM where
BreedName = a2.Breed2)
) a order by a.BreadName
Parameters...
Build Query...
Browse...
Parse Query
Preview...
OK
Cancel
Help

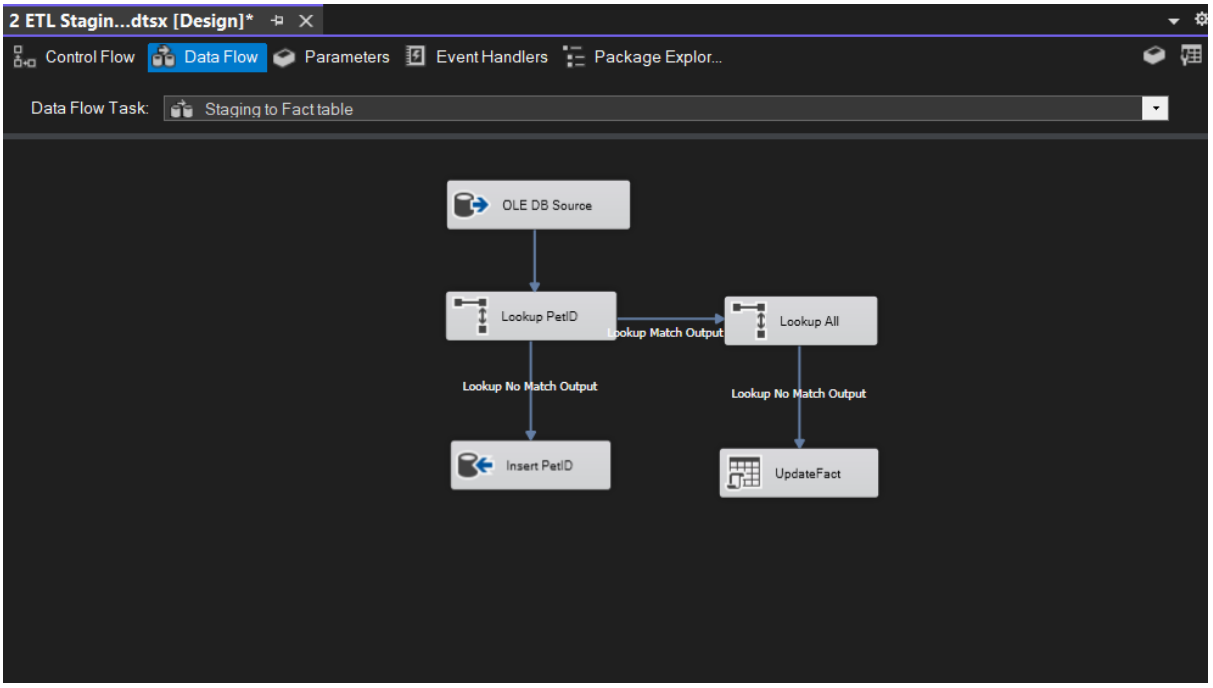
Sau khi thực hiện ETL từ Staging table vào Breed_Dim, chúng ta sẽ thực hiện update BreedID vào Staging table.



*Làm tương tự với các Dim còn lại (xem file SQL Command)

F. Xây dựng ETL cho Fact table

Lấy nguồn dữ liệu từ Staging table sau khi thực hiện các bước trên.
Dùng lookup để kiểm tra PetID có chưa
- Nếu chưa có ID thì insert vào Fact table.
- Nếu có rồi thì update các thông tin còn lại theo PetID.



G. Xây dựng ETL chạy song song

Tạo một Sequence Container chứa các ETL từ Staging table vào Dim table để chạy song song ETL dữ liệu.
Sau khi quá trình ETL thành công rồi mới tiến hành ETL từ staging table vào Fact table. (Như hình)