BÁO CÁO

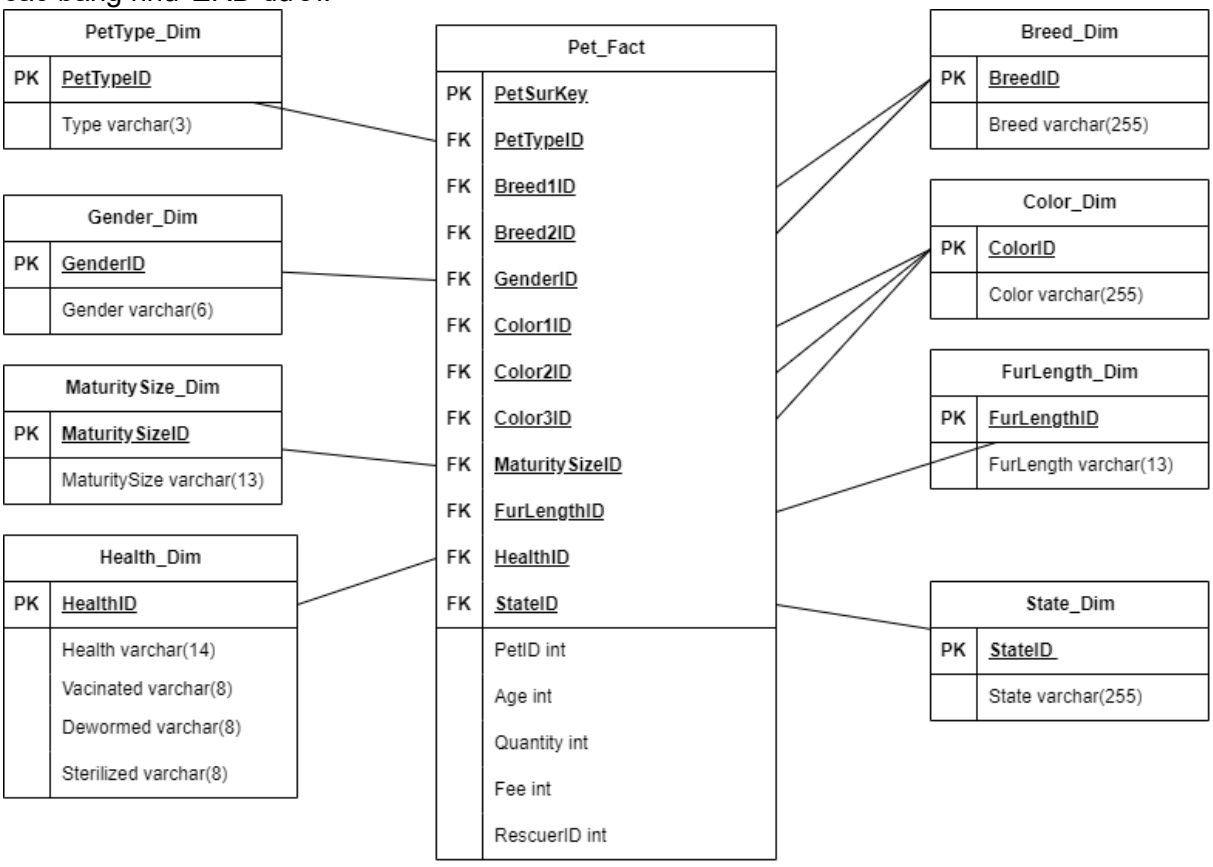
DEP302 – Assignment 01

Họ và tên: Nguyễn Hải Dương

Mã Sinh viên: FX16133

I. Thiết kế ERD

Data warehouse được thiết kế theo kiến trúc Centralized Data Warehouse. Bao gồm các bảng như ERD dưới:



Các câu lệnh SQL để xây dựng cơ sở dữ liệu
File: CreateTableQuery.sql

1. Tạo các bảng cho các dimensions.

```sql
--Create DIM TABLES
CREATE TABLE PetType_Dim (
    PetTypeID int IDENTITY(1,1) PRIMARY KEY,
    Type varchar(3) NOT NULL
);

CREATE TABLE Breed_Dim(
    BreedID int IDENTITY(1,1) PRIMARY KEY,
    Breed varchar(255) NOT NULL
);

CREATE TABLE Gender_Dim (
    GenderID int IDENTITY(1,1) PRIMARY KEY,
    Gender varchar(6) NOT NULL
);

CREATE TABLE Color_Dim (
    ColorID int IDENTITY(1,1) PRIMARY KEY,
    Color varchar(255) NOT NULL
);

CREATE TABLE MaturitySize_Dim (
    MaturitySizeID int IDENTITY(1,1) PRIMARY KEY,
    MaturitySize varchar(13) NOT NULL
);

CREATE TABLE FurLength_Dim (
    FurLengthID int IDENTITY(1,1) PRIMARY KEY,
    FurLength varchar(13) NOT NULL
);
CREATE TABLE Health_Dim (
    HealthID int IDENTITY(1,1) PRIMARY KEY,
    Health varchar(14) NOT NULL,
    Vacinated varchar(8) NOT NULL,
    Dewormed varchar(8) NOT NULL,
    Sterilized varchar(8) NOT NULL
);

CREATE TABLE State_Dim (
    StateID int IDENTITY(1,1) PRIMARY KEY,
    State varchar(255) NOT NULL
);
```

2. Tạo bảng cho fact table cùng các khóa phụ

```
-- CREATE FACT TABLE
GO
CREATE TABLE Pet_Fact (
    PetSurKey int IDENTITY(1,1) PRIMARY KEY,
    PetTypeID int NOT NULL,
    Breed1ID int NOT NULL,
    Breed2ID int NOT NULL,
    GenderID int NOT NULL,
    Color1ID int NOT NULL,
    Color2ID int NOT NULL,
    Color3ID int NOT NULL,
    MaturitySizeID int NOT NULL,
    FurlengthID int NOT NULL,
    HealthID int NOT NULL,
    StateID int NOT NULL,
    PetID int NOT NULL,
    Age int NOT NULL,
    Quantity INT NOT NULL,
    Fee int NOT NULL,
    RescuerID int NOT NULL,
    CONSTRAINT FK_Pet_Fact_Type_DIM FOREIGN KEY (PetTypeID)
        REFERENCES PetType_DIM (PetTypeID)
        ON DELETE CASCADE
        ON UPDATE CASCADE,
    CONSTRAINT FK_Pet_Fact_Breed1 FOREIGN KEY (Breed1ID)
        REFERENCES Breed_Dim (BreedID)
        ON DELETE CASCADE
        ON UPDATE CASCADE,
    CONSTRAINT FK_Pet_Fact_Breed2 FOREIGN KEY (Breed2ID)
        REFERENCES Breed_Dim (BreedID)
        ON DELETE NO ACTION
        ON UPDATE NO ACTION,
    CONSTRAINT FK_Pet_Fact_Gender FOREIGN KEY (GenderID)
        REFERENCES Gender_Dim (GenderID)
        ON DELETE CASCADE
        ON UPDATE CASCADE,
    CONSTRAINT FK_Pet_fact_Color1 FOREIGN KEY (Color1ID)
        REFERENCES Color_Dim (ColorID)
        ON DELETE CASCADE
        ON UPDATE CASCADE,
    CONSTRAINT FK_Pet_fact_Color2 FOREIGN KEY (Color2ID)
        REFERENCES Color_Dim (ColorID)
        ON DELETE NO ACTION
        ON UPDATE NO ACTION,
    CONSTRAINT FK_Pet_fact_Color3 FOREIGN KEY (Color3ID)
        REFERENCES Color_Dim (ColorID)
        ON DELETE NO ACTION
        ON UPDATE NO ACTION,
    CONSTRAINT FK_Pet_Fact_MaturitySize FOREIGN KEY (MaturitySizeID)
        REFERENCES MaturitySize_dim (MaturitySizeID)
        ON DELETE CASCADE
        ON UPDATE CASCADE,
    CONSTRAINT FK_Pet_Fact_FurLength FOREIGN KEY (FurLengthID)
        REFERENCES FurLength_Dim (FurLengthID)
        ON DELETE CASCADE
        ON UPDATE CASCADE,
    CONSTRAINT FK_Pet_Fact_Health FOREIGN KEY (HealthID)
        REFERENCES Health_Dim (HealthID)
        ON DELETE CASCADE
        ON UPDATE CASCADE,
    CONSTRAINT FK_Pet_Fact_State FOREIGN KEY (StateID)
        REFERENCES State_dim (StateID)
        ON DELETE CASCADE
        ON UPDATE CASCADE
)
```

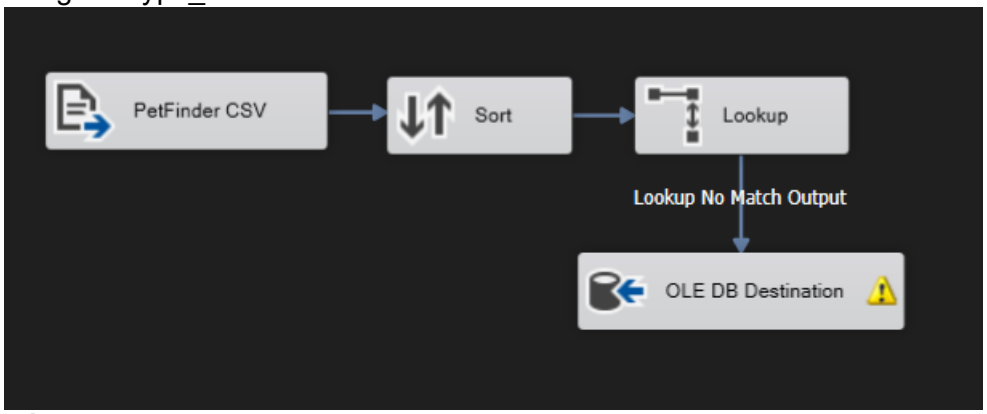
II. Xác định các truy vấn nghiệp vụ:

Các truy vấn nghiệp vụ được sử dụng cho cơ sở dữ liệu của bài

1. Số lượng thú nuôi chưa được tiêm vaccine và chưa được sổ giun, phân theo từng khu vực (state)
2. Số lượng thú cưng thuần chủng (Không lai giữa 2 giống khác nhau) phân thao từng loài
3. Phí nhận nuôi trung bình theo kích thước trưởng thành
4. Phí nhận nuôi trung bình theo giống (Bao gồm cả giống 1 và 2 nếu có)

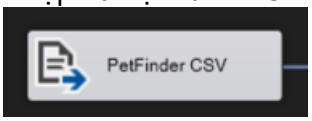
III. Xây dựng ETL cho Dimension tables:

1. Quy trình xây dựng ETL cho các Dimension table:
   Bảng PetType_Dimension

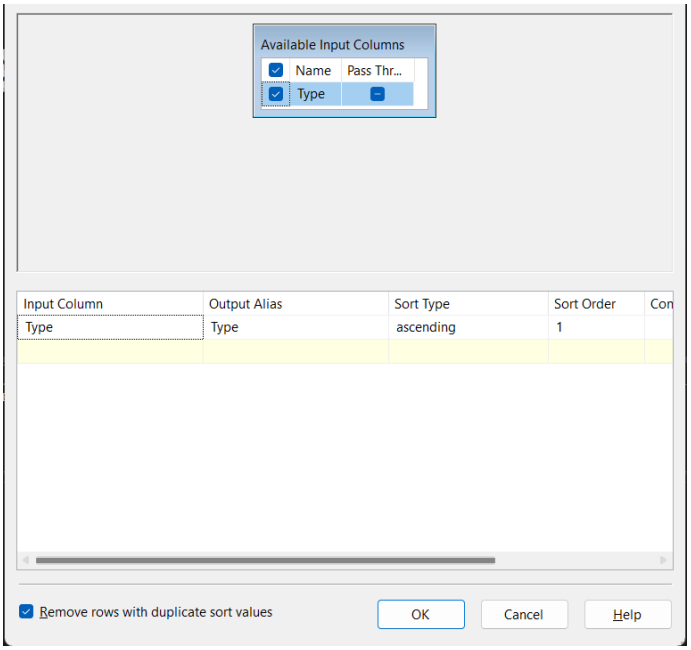


Gồm 4 bước:

a. Nhập dữ liệu từ file CSV

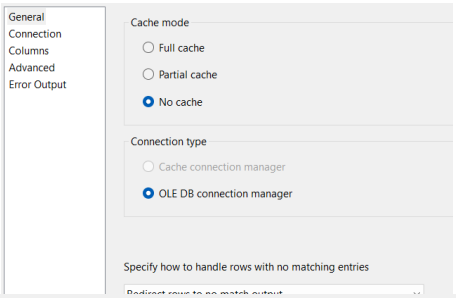


b. Sort dữ liệu theo type nhằm loại bỏ các giá trị bị trùng

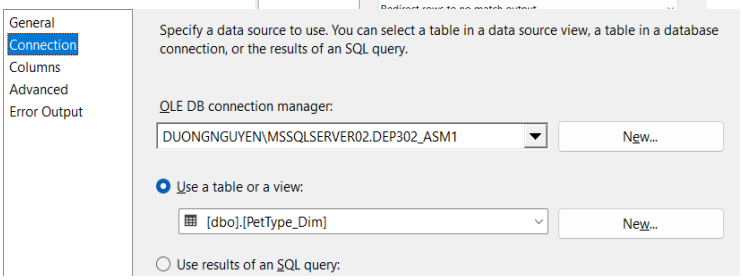


c. Lookup – Tìm/So sánh các dữ liệu đã có sẵn trong bảng PetType_Dim

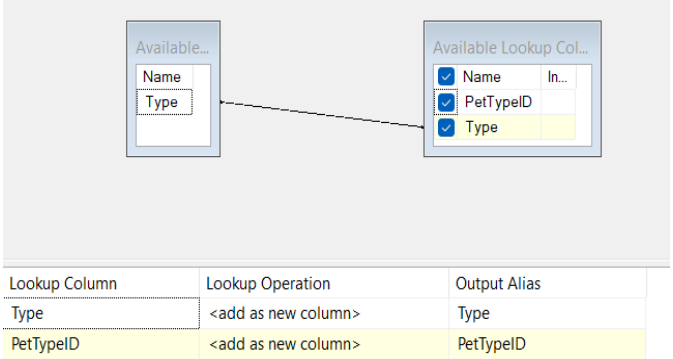
Chọn “Redirected rows to match output” để sử dụng những dữ liệu chưa có trong bảng PetType_Dim và nhập vào bảng này ở bước sau



Tạo Connection với bảng PetType_dim trong Database DEP302_ASM1



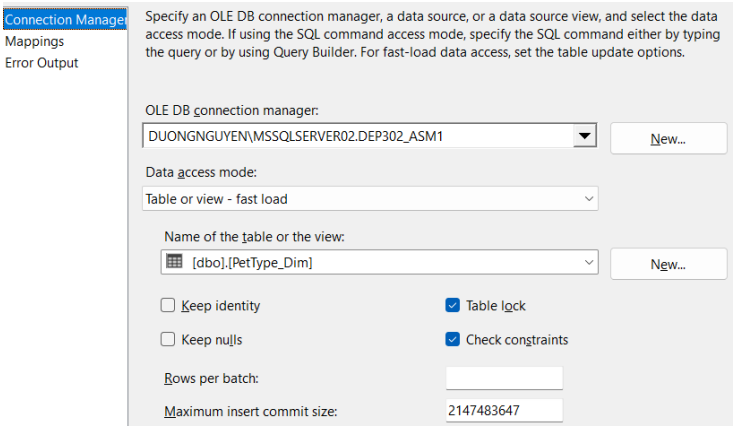
So sánh Type trong dữ liệu sau khi sort với dữ liệu look up trong bảng PetType_Dim



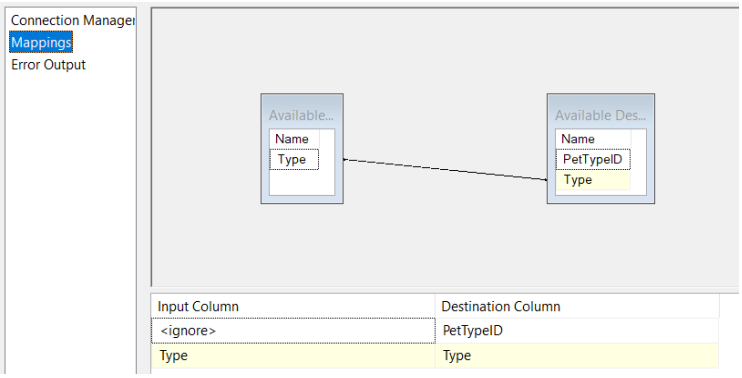
Kết nối dữ liệu của lookup theo dạng “Lookup no match output” với Destination

d. Truyền dữ liệu vào bảng PetType_Dim trong cơ sở dữ liệu:

Sử dụng OLE DB Destination để kết nối với bảng PetType_Dim



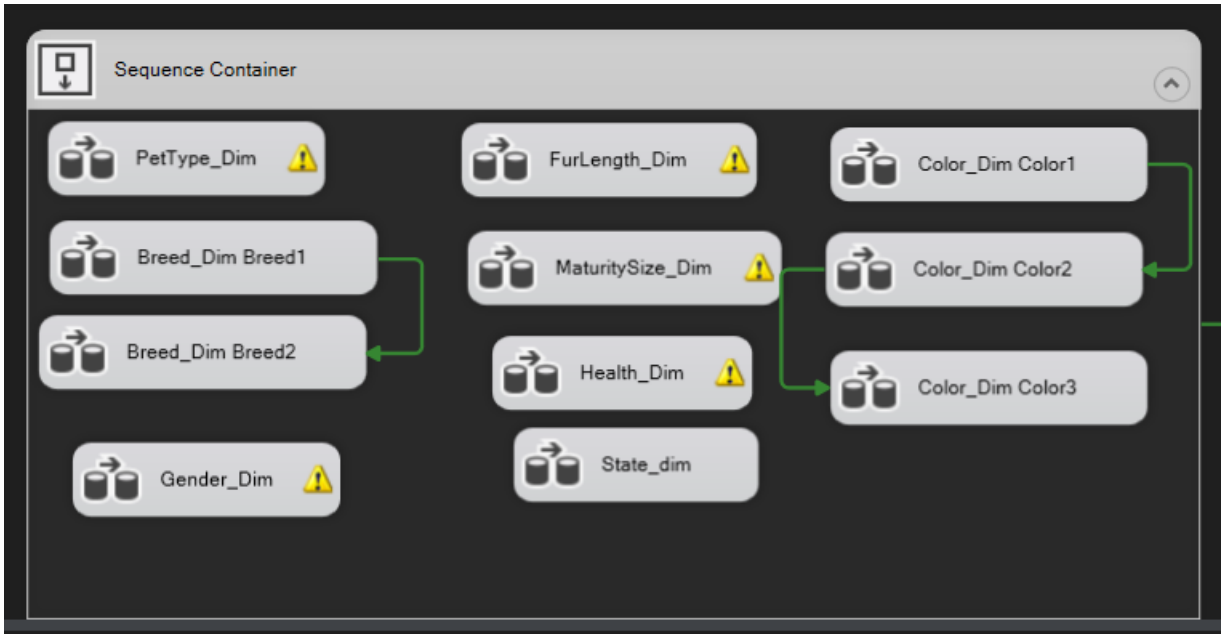
Mapping Type dữ liệu Type với Type trong bảng PetType_Dim



Tương tự với các dimension tables:

- FurLength_Dim
- Gender_Dim
- MaturittySize_Dim
- Health_Dim
- State_Dim
- Breed_Dim
- Color_Dim

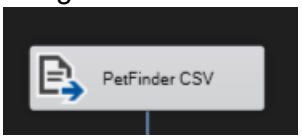
2. Thiết lập các Data Flow chạy đồng thời sử dụng Sequence Container:
Sau khi hoàn thành thiết lập cho các Dimension tables, ta được các các Data Flow như hình dưới.
Ta sử dụng Sequence Container để nhóm các Data Flow cho việc nhập dữ liệu các Dimension tables nhằm để cho các Data flow này được xử lý song song.



IV. Xây dựng ETL cho Fact table

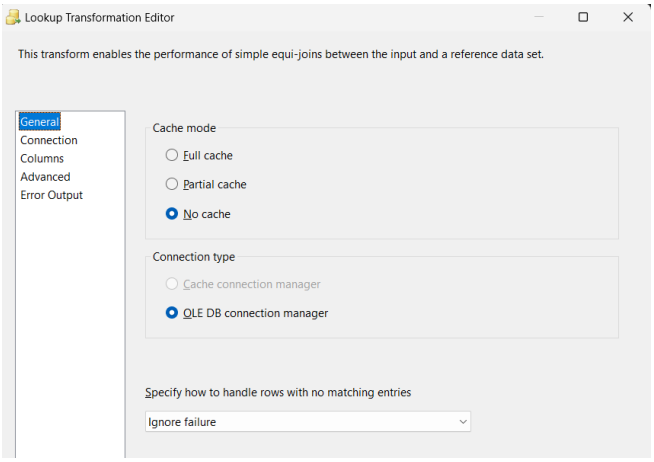
1. Tạo 1 Data Flow mới cho việc nhập dữ liệu Fact table:

2. Dùng Flat file source để làm nguồn dữ liệu:

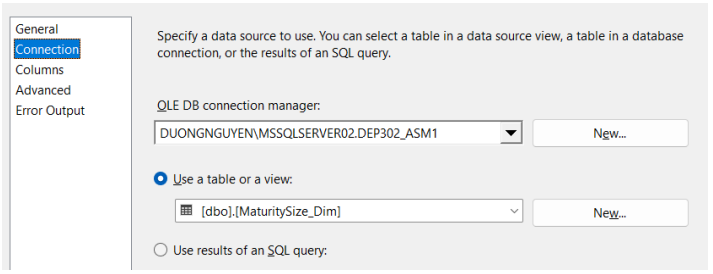


3. Sử dụng Lookup Transformation để tìm kiếm các dữ liệu liên quan:
   a. Thiết lập Lookup Transformation
   Ví dụ với bảng MaturitySize_Dim

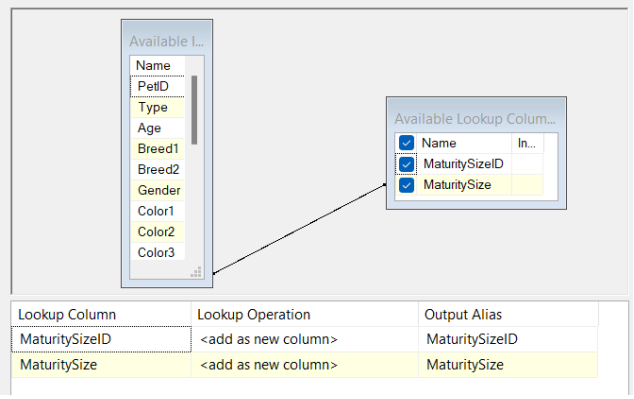
Thiết lập No Cache và kết nối với dữ liệu qua OLD DB Connection manager. Ignore các dữ liệu không match với thực thể



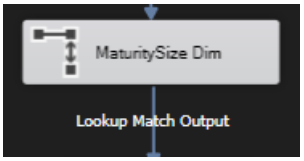
Kết nối với với cơ sở dữ liệu và chọn bảng MaturitySize_Dim



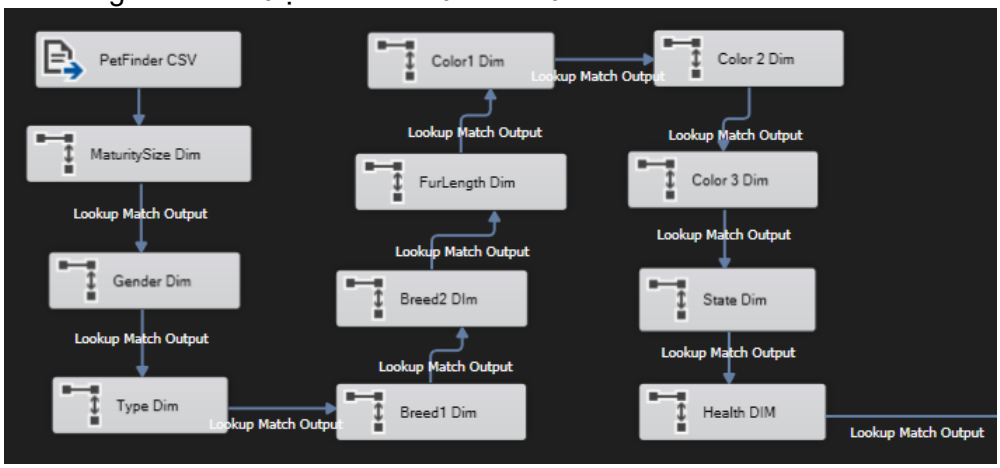
Lookup các dữ liệu có match với Maturity Size trong bảng MaturitySize_Dim
Và chọn sử dụng các trường MaturitySizeID và MaturitySize



Sử dụng output match với cài đặt Lookup Transformation ở trên để kết nối với hành động tiếp theo



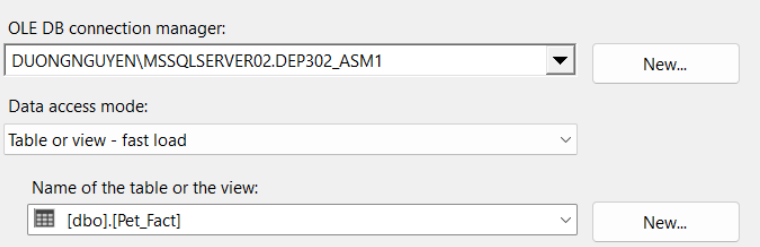
Tương tự, ta cài đặt Lookup Transformation cho các trường dữ liệu khác của bảng khác để được sơ đồ như hình dưới



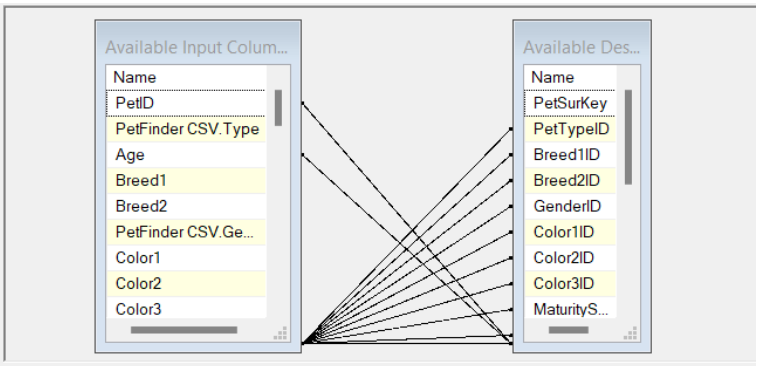
4. Nhập dữ liệu vào Fact Table

Sử dụng công cụ OLD DB Destination để nhập dữ liệu sau các Lookup Transformation

Kết nối với bảng Pet_Fact

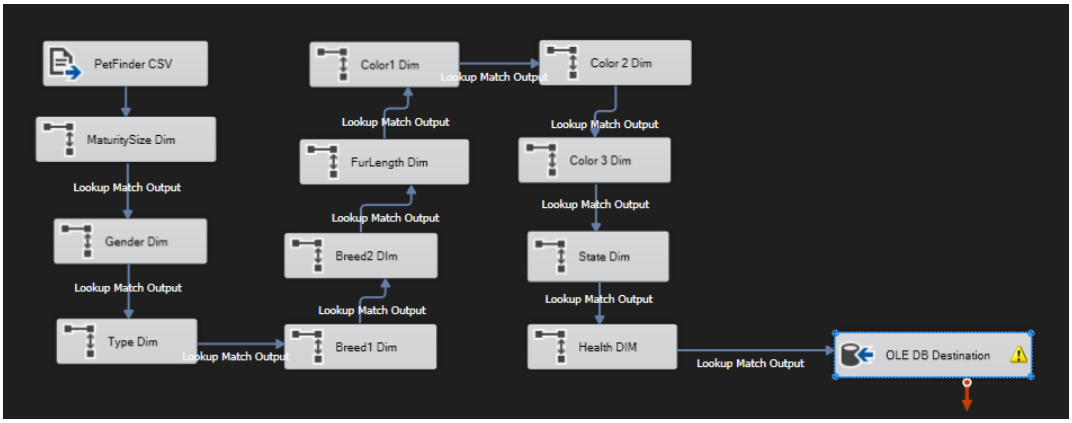


Mapping các dữ liệu có được từ output alias của các Lookup Transformation phù hợp với các cột của Fact Table



Input Column	Destination Column
<ignore>	PetSurKey
PetTypeID	PetTypeID
Breed1ID	Breed1ID
Breed2ID	Breed2ID
GenderID	GenderID
Color1ID	Color1ID
Color2ID	Color2ID
Color3ID	Color3ID
MaturitySizeID	MaturitySizeID
FurLengthID	FurlengthID

Sau khi hoàn thành, ta được Data Flow của Fact table như hình dưới:



V. Viết SQL để thực hiện các truy vấn nghiệp vụ

1. Số lượng thú nuôi chưa được tiêm vaccine và chưa được sổ giun, phân theo từng khu vực (state)

```
SELECT s.State, count(*) as TotalAnimalNeedCare
FROM Pet_fact p
JOIN Health_Dim h
ON h.HealthID = p.HealthID
JOIN state_dim s
ON p.stateID = s.stateID
WHERE h.vacinated != 'Yes'
    AND h.Dewormed != 'Yes'
Group by s.state
```

Results | Messages

	State	TotalAnimalNeedCare
1	Johor	1330
2	Kedah	315
3	Kelantan	40
4	Kuala Lumpur	8065
5	Labuan	5
6	Melaka	350
7	Negeri Sembilan	650
8	Pahang	195
9	Perak	710
10	Pulau Pinang	2130
11	Sabah	45
12	Sarawak	40
13	Selangor	17675
14	Terengganu	90

2. Số lượng từng loài thuần chủng (Không được lai giữa 2 loại khác nhau)

```
SELECT t.type, count(*) as totalPet
FROM Pet_Fact p
LEFT JOIN Breed_dim b
on p.Breed2ID = b.BreedID
LEFT JOIN PetType_Dim t
ON t.PetTypeID = p.PetTypeID
WHERE b.Breed = 'None'
GROUP BY t.Type
ORDER BY totalPet
```

	type	totalPet
1	Cat	25665
2	Dog	28130

3. Phí nhận nuôi theo trung bình theo kích thước trưởng thành

```sql
SELECT m.MaturitySize, AVG(p.fee) as AverageFee
FROM Pet_fact p
JOIN MaturitySize_Dim m
on p.MaturitySizeID = m.MaturitySizeID
GROUP BY MaturitySize
ORDER BY AverageFee DESC
```

Results Messages

	MaturitySize	AverageFee
1	Extra Large	59
2	Large	47
3	Small	23
4	Medium	17
5	Not Specified	2

4. Phí nhận nuôi trung bình theo giống

```sql
SELECT b.breed AS breed1, b2.breed AS breed2, AVG(p.fee) AS avgFee
FROM Pet_Fact p
JOIN Breed_Dim b ON p.Breed1ID = b.BreedID
JOIN Breed_Dim b2 ON p.Breed2ID = b2.BreedID
GROUP BY b.Breed, b2.Breed
ORDER BY avgFee DESC
```

Results Messages

	breed1	breed2	avgFee
1	Dilute Tortoiseshell	Domestic Long Hair	750
2	American Curl	Domestic Long Hair	650
3	Siberian Husky	German Shepherd Dog	600
4	German Shepherd Dog	Dachshund	550
5	Irish Setter	None	500
6	Jack Russell Terrier	Fox Terrier	500
7	American Bulldog	None	500
8	Jack Russell Terrier	Labrador Retriever	500
9	Shetland Sheepdog Sheltie	None	500
10	English Bulldog	None	470
11	West Highland White Terrier Westie	Poodle	400
12	Persian	Oriental Long Hair	400
13	Oriental Long Hair	Oriental Long Hair	400
14	German Shepherd Dog	Siberian Husky	400
15	Poodle	Schnauzer	380
16	American Curl	None	368
17	Siberian Husky	Siberian Husky	362

VI. Thiết kế các ETL chạy song song

Các ETL cho việc nhập dữ liệu vào các Dimension table được nhóm chung vào Sequence Contatiner để các ETL chạy song song.

Sau khi các ETL này hoàn thành thì việc nhập dữ liệu vào Fact table mới được thực thi.