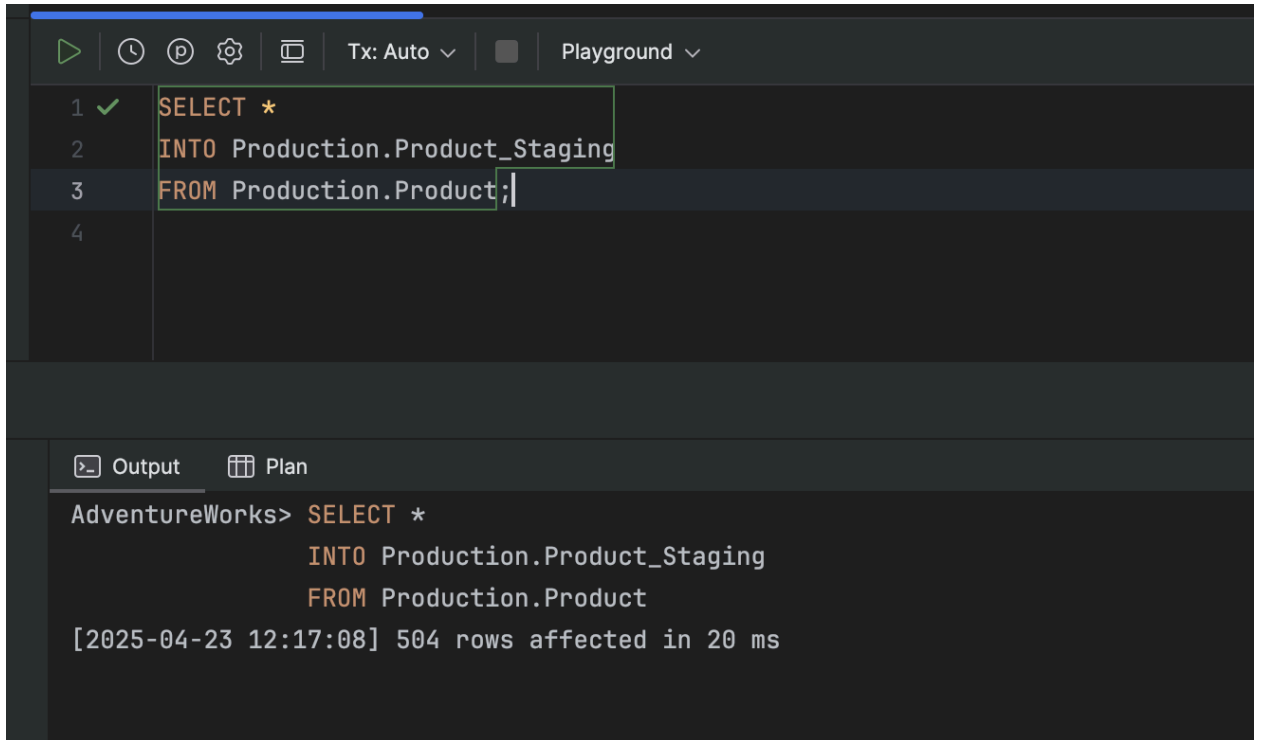


Veri Temizleme ve ETL Süreci Tasarımı

AdventureWorks veritabanındaki **Production.Product** tablosu üzerinde veri temizleme süreci tasarlandı. Bu süreçte, bazı veriler kasıtlı olarak bozulduktan sonra, yalnızca anlamlı değerlere sahip kayıtlar seçilerek yeni bir tabloya aktarıldı. Böylece hatalı veriler devre dışı bırakılmış oldu.

1. Staging Tablosunun Oluşturulması

İlk adım olarak, orijinal **Production.Product** tablosunun bir kopyası alındı. Böylece işlemler doğrudan ana tablo üzerinde değil, bu kopya üzerinde yapıldı.



```
1 ✓ SELECT *
2 INTO Production.Product_Staging
3 FROM Production.Product;
4
```

AdventureWorks> SELECT *
INTO Production.Product_Staging
FROM Production.Product
[2025-04-23 12:17:08] 504 rows affected in 20 ms

2. Veri Bozulması (Simülasyon Amaçlı)

Temizleme işlemlerinin uygulanabilmesi için, bazı veriler kasıtlı olarak geçersiz hale getirildi.

```
7 UPDATE Production.Product_Staging
8 SET SellStartDate = '1900-01-01'
9 WHERE ProductID % 4 = 0;
10

Output Plan
AdventureWorks> SELECT *
                INTO Production.Product_Staging
                FROM Production.Product
[2025-04-23 12:17:08] 504 rows affected in 20 ms
AdventureWorks> -- Fiyat alanı negatif değerlere çekildi
                UPDATE Production.Product_Staging
                SET ListPrice = -99.99
                WHERE ProductID % 10 = 0;

                -- Satış başlangıç tarihi anlamını yitirecek şekilde geriye alındı
                UPDATE Production.Product_Staging
                SET SellStartDate = '1900-01-01'
                WHERE ProductID % 4 = 0;
[2025-04-23 12:17:49] 177 rows affected in 43 ms
```

Bu işlemler sonucunda, belirli satırlardaki fiyatlar negatif hale geldi ve bazı tarihler anlamsız derecede eskiye çekildi.

3. Verilerin Temizlenmesi (Bozuk Verilerin Dışlanması)

Temiz bir veri kümesi elde etmek amacıyla, yalnızca anlamlı fiyat ve tarih bilgilerine sahip kayıtlar seçildi.

```
1 ✓ SELECT *
2 INTO Production.Product_Cleaned
3 FROM Production.Product_Staging
4 WHERE ListPrice >= 0
5 AND SellStartDate >= '2000-01-01';
6
```

Output Plan

```
AdventureWorks> SELECT *
                  INTO Production.Product_Cleaned
                  FROM Production.Product_Staging
                  WHERE ListPrice >= 0
                        AND SellStartDate >= '2000-01-01'
[2025-04-23 12:19:20] 351 rows affected in 41 ms
```

4. Temizlik İşleminin Değerlendirilmesi

Temizleme işleminden sonra, kaç kaydın devre dışı bırakıldığını görmek için bir özet sorgu çalıştırıldı:

```

1 ✓ SELECT
2     COUNT(*) AS TotalRecords,
3     SUM(CASE WHEN ListPrice < 0 THEN 1 ELSE 0 END) AS BozukFiyatSayisi,
4     SUM(CASE WHEN SellStartDate < '2000-01-01' THEN 1 ELSE 0 END) AS GecersizTarihSayisi
5 FROM Production.Product_Staging;
6

```

Output | Plan

Result 453

×

Table Icon

Chart Icon

Refresh Icon

Undo Icon

Stop Icon

Filter Icon

Search Icon

Download Icon

	TotalRecords	BozukFiyatSayisi	GecersizTarihSayisi
1	504	50	127

1 ✓

2

3

4

5

6

```
SELECT
COUNT(*) AS TotalRecords,
SUM(CASE WHEN ListPrice < 0 THEN 1 ELSE 0 END) AS BozukFiyatSayisi,
SUM(CASE WHEN SellStartDate < '2000-01-01' THEN 1 ELSE 0 END) AS GecersizTarihSayisi
FROM Production.Product_Cleaned;
```

Output | Plan

Result 454 ×

TotalRecords ▾

BozukFiyatSayisi ▾

GecersizTarihSayisi ▾

1

351

0

0

Bu sonuçlar, veri temizliğinin etkisini ve ne kadar kayıt dışlandıldığını sayısal olarak göstermektedir.

Farklı Kaynaklardan Gelen Verilerin Dönüştürülmesi ve Birleştirilmesi

Bu aşamada, iki farklı veri kaynağından gelen personel bilgileri üzerinde veri dönüştürme işlemleri gerçekleştirilmiş ve veriler ortak bir formata getirilerek tek bir yapıda birleştirilmiştir. Bu

işlem, sistemler arası veri entegrasyonu örneği olarak tasarlanmış ve ETL sürecinin uygulamalı bir adımı olarak gerçekleştirilmiştir.

Auto ▾

Playground ▾

Import

+

−

📄

⬆

⬇

⬆

⬇

↩

➡

from: **personel_A.csv**

to: **personel_A** master.dbo [localhost]

to: **personel_A**

Schema **dbo master [localhost]**

Table **personel_A** New: auto >

Mapping

to: Column	to: Type	from: Column
ad_soyad	VARCHAR(MAX)	ad_soyad
yas	int	<Auto> (yas)
cinsiyet	VARCHAR(MAX)	cinsiyet
maas	int	<Auto> (maas)

☒ Write errors to file /private/var/folders/q5/y3y5nv8j4j7gvdq74j8qv

☐ Insert inconvertible values as null

☐ Disable indexes and triggers, lock table (may be faster)

Data Preview

DDL Preview

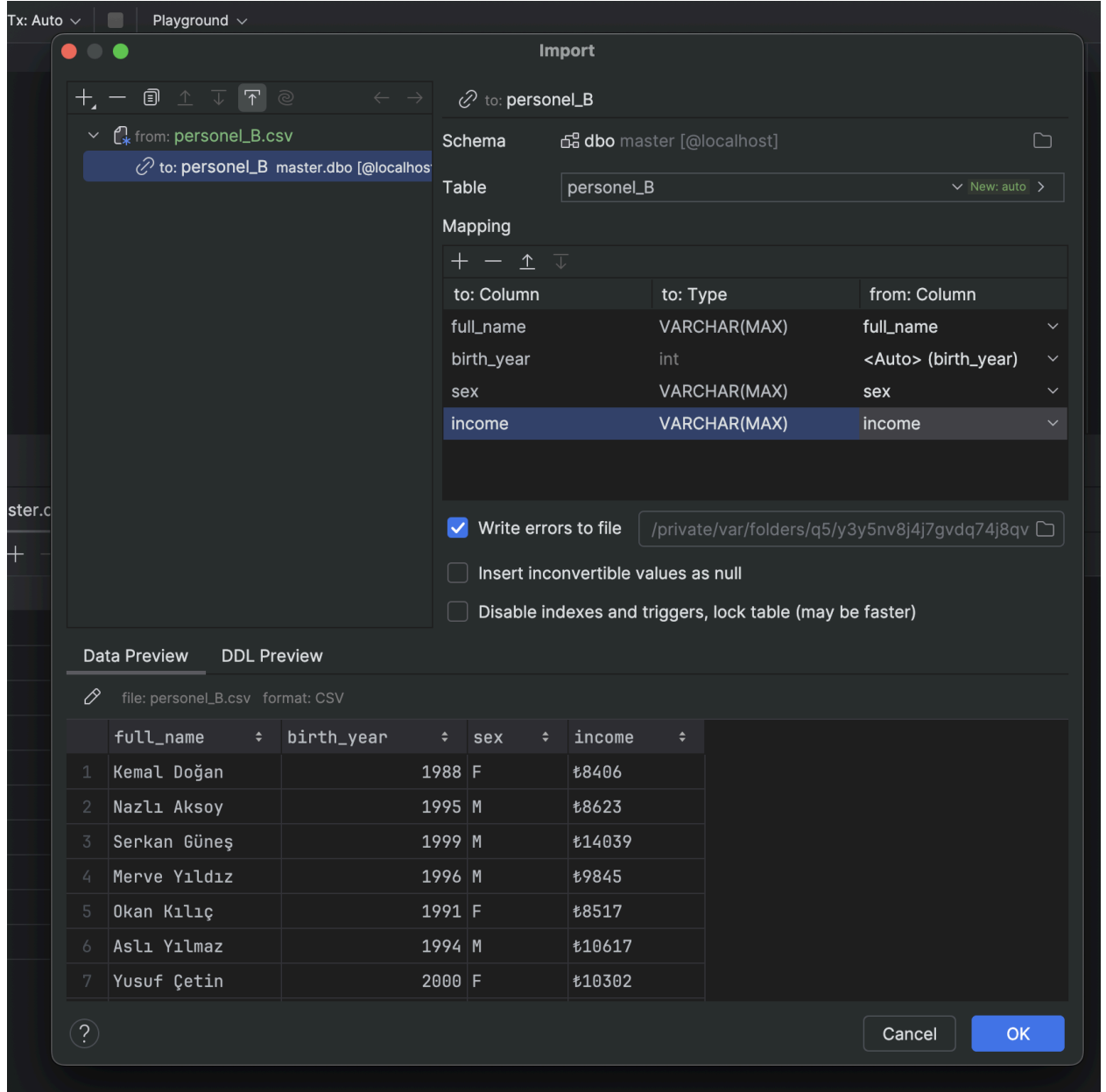
file: personel_A.csv format: CSV

	ad_soyad	yas	cinsiyet	maas
1	Ali Yılmaz	29	erkek	11528
2	Ayşe Demir	39	erkek	14012
3	Mehmet Koç	31	kadın	11792
4	Elif Yalçın	37	kadın	8919
5	Can Kaya	35	kadın	9947
6	Zeynep Şahin	33	erkek	10763
7	Burak Arslan	26	kadın	10844

?

Cancel

OK



İki farklı sistemden gelen veriler personel_A ve personel_B tablolarına aktarıldı.

Verilerin dönüştürülmesi

İki tablo arasındaki kolon isimleri ve veri tipleri farklı olduğundan, veriler ortak bir yapıya getirilmeden önce dönüştürme işlemleri uygulanmıştır. Amaç, iki kaynaktan gelen bilgilerin tek bir birleşik tablo altında birleştirilmesini mümkün kılmaktır.

```
1 ✓ SELECT
2     ad_soyad AS name,
3     yas AS age,
4     CASE
5         WHEN cinsiyet = 'erkek' THEN 'Male'
6         WHEN cinsiyet = 'kadın' THEN 'Female'
7         ELSE 'Other'
8     END AS gender,
9     maas AS salary
10 INTO personel_cleaned
11 FROM personel_a;
12
```

```
1 ✓ INSERT INTO personel_cleaned (name, age, gender, salary)
2 SELECT
3     full_name AS name,
4     YEAR(GETDATE()) - birth_year AS age,
5     CASE
6         WHEN sex = 'M' THEN 'Male'
7         WHEN sex = 'F' THEN 'Female'
8         ELSE 'Other'
9     END AS gender,
10    REPLACE(REPLACE(income, '₺', ''), '.', '') AS salary
11 FROM personel_b;
12
```

Bu adım sonucunda, farklı formatlara ve veri yapılarına sahip iki kaynaktan alınan veriler tek bir biçime dönüştürülmüş ve analiz için uygun olan **personel_cleaned** adlı tabloya aktarılmıştır.


```
1 ✓ SELECT * FROM personel_cleaned;
```



Output | Plan **personel_cleaned** ×



	name ▾	age ▾	gender ▾	salary ▾
1	Ali Yilmaz	29	Male	11528
2	Ayşe Demir	39	Male	14012
3	Mehmet Koç	31	Female	11792
4	Elif Yalçın	37	Female	8919
5	Can Kaya	35	Female	9947
6	Zeynep Şahin	33	Male	10763
7	Burak Arslan	26	Female	10844
8	Sena Kurt	41	Male	8077
9	Ahmet Tunç	40	Female	9585
10	Fatma Aydın	37	Female	13806
11	Kemal Doğan	37	Female	8406
12	Nazlı Aksoy	30	Male	8623
13	Serkan Güneş	26	Male	14039
14	Merve Yıldız	29	Male	9845
15	Okan Kiliç	34	Female	8517
16	Aslı Yılmaz	31	Male	10617
17	Yusuf Çetin	25	Female	10302
18	Buse Korkmaz	41	Male	10877
19	Cemre Öztürk	41	Female	9816
20	Eren Dursun	42	Male	11283