

Database Normalization

Database normalizasyonu bir tablodaki(data) tekrarları azaltmak için o tabloyu birden fazla tabloya ayırmayı ifade eder.

Elimizde aşağıdaki gibi bazı kitapların isimleri, yazarları, türleri ve basım yıllarına ilişkin bir tablo olduğunu düşünelim:

| Book Title | Authors | Genre | City | Country |
|-----------------------|---------------------|--------------------|--------------|---------|
| The Hobbit | J.R.R. Tolkien | Fantasy | London | UK |
| 1984 | George Orwell | Dystopian, Classic | London | UK |
| Animal Farm | George Orwell | Dystopian | London | UK |
| Pride and Prejudice | Jane Austen | Romance | London | UK |
| Emma | Jane Austen | Romance, Gothic | London | UK |
| To Kill a Mockingbird | Harper Lee | Classic | Philadelphia | USA |
| The Great Gatsby | F. Scott Fitzgerald | Classic | New York | USA |
| Wuthering Heights | Emily Bronte | Gothic, Romance | London | UK |
| Jane Eyre | Charlotte Bronte | Gothic | London | UK |
| Fahrenheit 451 | Ray Bradbury | Dystopian | New York | USA |
| The Hobbit | J.R.R. Tolkien | Fantasy | London | UK |
| The Hobbit | J.R.R. Tolkien | Fantasy | London | UK |
| Emma | Jane Austen | Romance, Gothic | London | UK |

Öncelikle bu tablodaki problem nedir?

- Tekrarlı satırlar (duplicated rows) var. The Hobbit kitabı ile ilgili tekrarlı satırlarımız var.
- Bazı yazar isimleri (Tolkien, Orwell, Austen) ve kitap türlerinde tekrarlar var.
- Bazı sütunlarda değerler birden çok tekrar ediyor. 2., 5., 9., ve son satırda genre sütunu için birden fazla değer var.

Bu problemlerden arındırmamız gerekiyor datayı.

Tekrarlı satırları silmenin mantığı çok basit. Bize faydalı değiller. Bir tane satır yeterli.

Yazar tekrarlarını ise istemiyoruz. Örneğin verimizin yanlış olduğunu düşünelim: George Orwell'in ikinci bir ismi olduğunu öğrendik ve bunu datamızda güncellememiz gerek. Tek tek George Orwell tarafından yazılmış tüm kitaplar için satırlara gidip güncelleme yapmamız gerekecekti. Bu pek problem değilmiş gibi gözüküyor: 15 satırımız var sadece. Ancak farklı bir data düşünelim:

- Amazon'da bir satıcıyız. Datamız var: bu datamızda sattığımız ürün, müşterimizin adı, adresi vs bilgiler var. Bir müşterimizin adresini değiştirdiğini düşünelim. 10000 satırlık veride o müşteriye ait tüm satışlar için adres bilgisini güncellememiz gerekecekti.

Yine genre kısmında birden fazla değer olması istediğimiz bir durum değil.

Bu tabloyu normalize ederek bu uzun işlemlerden kurtulabiliriz.

Sırasıyla aşağıdaki adımları takip ederek bu datayı birden çok tabloya ayırabiliriz. 3 adımdan oluşuyor normalizasyonumuz.

1.Adım: 1NF – First Normal Form

Bu ilk adımımız. Verinin ilk değişimini ifade ediyor (first normal form).

1NF de aşağıdaki şartlar sağlanmalı:

- *Tekrarlı satırları sil. Tekrarlı satır tüm sütun değerlerinin aynı olduğu satırları ifade eder.* Bizde The hobbit ve 1984 sütunları tekrarlanmış. Onları siliyoruz.
- *Her sütunda tek bir değer olmasını sağla. Bu her satır için tüm sütunlarda tek bir değer olmalı.* Bizim datamız zaten böyle ancak mesela 1984 kitabının genre sütununda. Bu durumu çözmek için birden fazla olan değerlerin her biri için yeni bir satır yaratıyoruz.

Bu adımları uyguladıktan sonra verimiz şöyle gözüküyor:

| Book Title | Authors | Genre | City | Country |
|-----------------------|---------------------|-----------|--------------|---------|
| The Hobbit | J.R.R. Tolkien | Fantasy | London | UK |
| 1984 | George Orwell | Dystopian | London | UK |
| 1984 | George Orwell | Classic | London | UK |
| Animal Farm | George Orwell | Dystopian | London | UK |
| Pride and Prejudice | Jane Austen | Romance | London | UK |
| Emma | Jane Austen | Romance | London | UK |
| Emma | Jane Austen | Gothic | London | UK |
| To Kill a Mockingbird | Harper Lee | Classic | Philadelphia | USA |
| The Great Gatsby | F. Scott Fitzgerald | Classic | New York | USA |
| Wuthering Heights | Emily Bronte | Gothic | London | UK |
| Wuthering Heights | Emily Bronte | Romance | London | UK |
| Jane Eyre | Charlotte Bronte | Gothic | London | UK |
| Fahrenheit 451 | Ray Bradbury | Dystopian | New York | USA |

Verimiz 1NF şartlarına uyuyor. 2NF ye geçebiliriz.

Özet: birinci adımımız (1NF) tekrar eden satırları ve birden fazla değere sahip hücreleri sildi. Ancak hala devam eden bir problem var: aynı bilgiler sürekli tekrar ediyor ve çok yer kaplıyor.

2.Adım: 2NF Second Normal Form

Bu adım biraz daha karışık. Uygulayabilmek için şu kavramı bilmek gerekiyor: **Functional dependency**.

Functional dependency

Veri setinde sütunları birbirleri ile ilişkili olma durumu. Yani bir sütunun diğer sütunun değerini belirlemesi durumu.

Matematiksel düşünelim:

$x = y + 2$ denkleminde x y nin değerlerine bağlıdır. Yani x y'ye functional dependent'tır. y nin her bir değeri için belirli tek bir x değeri mevcuttur.

Örnek veri setimizde bu konuyu her sütun için düşünelim:

- Book title Book title herhangi bir sütuna bağlı mı? Tek tek diğer satırlar için kontrol edelim. Öncelikle book title ve authors için bakalım: bir kitabın ismi kitabın yazarını belirleyebilir mi? Hayır. Çünkü aynı isimde bir kitap farklı yazarlar tarafından yazılabilir. Bu durumda Book title ve yazarlar arasında bir functional dependency yok. Book title ve genre arasında var mı? Bir kitabın ismi genresini belirleyebilir mi? Hayır. Tüm sütunlar için de bu durum böyle. Yani book title için herhangi bir dependency yok.
- Authors Bir kitabın yazarı kitabın ismini belirleyebilir mi? Yani bir kitabın yazarının adı George Orwell olduğunda örneğin kitabın ismi tek bir değer mi alır? Hayır. Geor Orwell birden fazla

kitap yazmış olabilir. Veya aynı kitap ismi başka bir yazar tarafından yazılmış olabilir. Dolayısıyla bir functional dependency yoktur.

- City City book title ı belirleyebilir mi? Hayır. Author'ı? Hayır. Peki Country'yi? Evet!. City London olduğunda country tek değer alabilir: UK. Başka bir değer alamaz. Yani city ile country arasında bir functional dependency var. Country city'ye bağlı. Peki tersi doğru mu? Yani city country'ye bağlı mı? Hayır. Çünkü country UK olduğunda city Liverpool olabilir mesela. Dolayısıyla bir bağımlılık yok.

Diğer sütunlar için baktığımızda verimizde tek functional dependency var. City ve Country arasında.

Başka bir örnek. Aşağıdaki tabloya bakın:

| Student No | Course No | Course Fee |
|------------|-----------|------------|
| 1 | c1 | 1000 |
| 2 | c2 | 1500 |
| 1 | c4 | 2000 |
| 4 | c3 | 1000 |
| 4 | c1 | 1000 |
| 2 | c5 | 2000 |

Burada functional dependency nerede?

Student no Course no yu belirleyebilir mi? Yani her bir student no için tek bir course no mu var? Hayır. Student No 1 için c1 ve c4 var örneğin. Çünkü bir öğrenci birden fazla kurs alabilir.

Ancak course fee course no ya bağlıdır. Çünkü her bir course no için tek bir course fee değeri vardır.

2.Adım: 2NF Second Normal Form

Functional dependencyyi anladıktan sonra 2.adıma geçebiliriz. Bu adımda artık datayı farklı tablolara ayıracağız.

Ayrırma işlemini functional dependency'ye göre yapacağız. Tablomuza yeniden bakalım:

| Book title | Authors | Genre | City | Country |
|-----------------------|---------------------|-----------|--------------|---------|
| The Hobbit | J.R.R. Tolkien | Fantasy | London | UK |
| 1984 | George Orwell | Dystopian | London | UK |
| 1984 | George Orwell | Classic | London | UK |
| Animal Farm | George Orwell | Dystopian | London | UK |
| Pride and Prejudice | Jane Austen | Romance | London | UK |
| Emma | Jane Austen | Romance | London | UK |
| Emma | Jane Austen | Gothic | London | UK |
| To Kill a Mockingbird | Harper Lee | Classic | Philadelphia | USA |
| The Great Gatsby | F. Scott Fitzgerald | Classic | New York | USA |
| Wuthering Heights | Emily Bronte | Gothic | London | UK |
| Wuthering Heights | Emily Bronte | Romance | London | UK |
| Jane Eyre | Charlotte Bronte | Gothic | London | UK |
| Fahrenheit 451 | Ray Bradbury | Dystopian | New York | USA |

Tek functional dependency city ve country arasındaydı. Bunları ayıracağız Ve tek bir tablo yapacağız. Yani bir tablomuz kesinlikle belirli: City ve Country.

Öncelikle bu tablomuzu yapalım:

City'ye bakıyoruz. Tekil değerleri(unique values) alıyoruz: London, Philadelphia, New York. Bunların karşılıklarına Countrylerini yazıyoruz.

| City | Country |
|--------------|---------|
| London | UK |
| Philadelphia | USA |
| New York | USA |

Bu işlemin ardından bir id tanımlıyoruz. **LocationID** diyelim LocationID her satır için tek bir değer alan bir id. Genelde 1,2,3... diye gider. Onu da tablonun başına sütun olarak koyuyuz. Artık **LocationID** bu tablomuz için Primary Key olmuş oldu. Çünkü her bir **LocationID** için farklı city ve country değerleri var:

| LocationID | City | Country |
|------------|--------------|---------|
| 1 | London | UK |
| 2 | Philadelphia | USA |
| 3 | New York | USA |

Bu tablomuz hazır. Sadece isim vermek kaldı. **Locations** diyelim tablonun adına.

Peki ayırma işlemimiz bitti mi? Bu tabloyu yaratınca ana tablomuzdan bu sütunları çıkarmış olduk. Peki bakalım mevcut haline ana tablomuzun:

| Book title | Authors | Genre |
|-----------------------|---------------------|-----------|
| The Hobbit | J.R.R. Tolkien | Fantasy |
| 1984 | George Orwell | Dystopian |
| 1984 | George Orwell | Classic |
| Animal Farm | George Orwell | Dystopian |
| Pride and Prejudice | Jane Austen | Romance |
| Emma | Jane Austen | Romance |
| Emma | Jane Austen | Gothic |
| To Kill a Mockingbird | Harper Lee | Classic |
| The Great Gatsby | F. Scott Fitzgerald | Classic |
| Wuthering Heights | Emily Bronte | Gothic |
| Wuthering Heights | Emily Bronte | Romance |
| Jane Eyre | Charlotte Bronte | Gothic |
| Fahrenheit 451 | Ray Bradbury | Dystopian |

Hala tekrarlar var. Ancak functional dependency yok bu sütunlar arasında. Nasıl ayıracağız tekrarlardan kurtulmak için? Artık functional dependency kalmadığı için her bir sütunun kendi tablosunu oluşturacağız. Book title ı en son yapalım. Bunun nedeni kısaca şu: data kitaplarla ilgili olduğu için book title ana tablomuzda olacak.

Dolayısıyla Authorsdan devam edelim:

Sırayla gidelim. Her bir tekil yazar adını alıp tabloya koyuyoruz. Ardından **AuthorID** oluşturuyoruz.

| AuthorID | Authors |
|----------|---------------------|
| 1 | J.R.R. Tolkien |
| 2 | George Orwell |
| 3 | Jane Austen |
| 4 | Harper Lee |
| 5 | F. Scott Fitzgerald |
| 6 | Emily Bronte |
| 7 | Charlotte Bronte |
| 8 | Ray Bradbury |

Tablomuzun adı Authors olsun.

Genres için yapalım:

| GenreID | Genre |
|---------|-----------|
| 1 | Fantasy |
| 2 | Dystopian |
| 3 | Classic |
| 4 | Romance |
| 5 | Gothic |

Genres olsun tablomuzun adı.

Şimdi aynı işlemi ana tablomuz için yapalım:

| BookID | Book Title |
|--------|-----------------------|
| 1 | The Hobbit |
| 2 | 1984 |
| 3 | Animal Farm |
| 4 | Pride and Prejudice |
| 5 | Emma |
| 6 | To Kill a Mockingbird |
| 7 | The Great Gatsby |
| 8 | Wuthering Heights |
| 9 | Jane Eyre |
| 10 | Fahrenheit 451 |

Books olsun tablomuzun adı.

Tüm tablolarımız hazır. Burada bitirebiliriz!

Ancak ek olarak şöyle bir tablo yapabiliriz:

Books tablomuzla diğer tabloları birleştirmek. Neden books tablosuyla? Çünkü konumuz bookslarla ilgili. Diğer tablolarla booksu da birleştirebilirdik ama mantıklı olan booksa diğer tabloları dahil etmek.

Nasıl dahil edeceğiz? Çok basit. İlk tablomuzla bir daha bakalım(1NF'deki):

| Book title | Authors | Genre | City | Country |
|-----------------------|---------------------|-----------|--------------|---------|
| The Hobbit | J.R.R. Tolkien | Fantasy | London | UK |
| 1984 | George Orwell | Dystopian | London | UK |
| 1984 | George Orwell | Classic | London | UK |
| Animal Farm | George Orwell | Dystopian | London | UK |
| Pride and Prejudice | Jane Austen | Romance | London | UK |
| Emma | Jane Austen | Romance | London | UK |
| Emma | Jane Austen | Gothic | London | UK |
| To Kill a Mockingbird | Harper Lee | Classic | Philadelphia | USA |
| The Great Gatsby | F. Scott Fitzgerald | Classic | New York | USA |
| Wuthering Heights | Emily Bronte | Gothic | London | UK |
| Wuthering Heights | Emily Bronte | Romance | London | UK |
| Jane Eyre | Charlotte Bronte | Gothic | London | UK |
| Fahrenheit 451 | Ray Bradbury | Dystopian | New York | USA |

Burada sırayla şu işlemleri yapıyoruz:

- Book title için birleştirme yaptığımız için ona dokunmuyoruz. Her bir Book title için BookID değerini ekliyoruz.

| BookID | Book title | Authors | Genre | City | Country |
|--------|-----------------------|---------------------|-----------|--------------|---------|
| 1 | The Hobbit | J.R.R. Tolkien | Fantasy | London | UK |
| 2 | 1984 | George Orwell | Dystopian | London | UK |
| 2 | 1984 | George Orwell | Classic | London | UK |
| 3 | Animal Farm | George Orwell | Dystopian | London | UK |
| 4 | Pride and Prejudice | Jane Austen | Romance | London | UK |
| 5 | Emma | Jane Austen | Romance | London | UK |
| 5 | Emma | Jane Austen | Gothic | London | UK |
| 6 | To Kill a Mockingbird | Harper Lee | Classic | Philadelphia | USA |
| 7 | The Great Gatsby | F. Scott Fitzgerald | Classic | New York | USA |
| 8 | Wuthering Heights | Emily Bronte | Gothic | London | UK |
| 8 | Wuthering Heights | Emily Bronte | Romance | London | UK |
| 9 | Jane Eyre | Charlotte Bronte | Gothic | London | UK |
| 10 | Fahrenheit 451 | Ray Bradbury | Dystopian | New York | USA |

Bu işlemin ardından her bir sütun için onların ID lerini koyuyoruz.

| BookID | Book Title | AuthorID | GenreID | LocationID |
|--------|-----------------------|----------|---------|------------|
| 1 | The Hobbit | 1 | 1 | 1 |
| 2 | 1984 | 2 | 2 | 1 |
| 3 | 1984 | 2 | 3 | 1 |
| 4 | Animal Farm | 2 | 2 | 1 |
| 5 | Pride and Prejudice | 3 | 4 | 1 |
| 6 | Emma | 3 | 4 | 1 |
| 7 | Emma | 3 | 5 | 1 |
| 8 | To Kill a Mockingbird | 4 | 3 | 2 |
| 9 | The Great Gatsby | 5 | 3 | 3 |
| 10 | Wuthering Heights | 6 | 5 | 1 |

| BookID | Book Title | AuthorID | GenreID | LocationID |
|--------|-------------------|----------|---------|------------|
| 11 | Wuthering Heights | 6 | 4 | 1 |
| 12 | Jane Eyre | 7 | 5 | 1 |
| 13 | Fahrenheit 451 | 8 | 2 | 3 |

bu tablo yukarıdakinin devamı.
Ayrı bir tablo değil.

Bu tablo da artık hazır durumda.

Peki neden **Book Title** ı tuttuk? Çünkü verimizin konusu için önemli. Mevcut durumda **AuthorID**, **GenreID**, ve **LocationID** foreign key. Neden? Çünkü bu sütunlar başka tablolardan geliyor. **BookID** bu veri seti için Primary Key. Neden? Çünkü her bir satır için tek değer alıyor.

Example

- “Thriller” by Michael Jackson, released in 1982, is one of the best-selling albums of all time, spanning genres like pop, post-disco, rock, and funk.
- “Back in Black” by AC/DC, a hard rock album, was released in 1980 as a tribute to their former vocalist.
- Madonna’s “Like a Virgin,” released in 1984, is a pop and dance album that became an international success.
- “The Dark Side of the Moon” by Pink Floyd, released in 1973, is a progressive rock album known for its philosophical lyrics and experimental sound.
- “Led Zeppelin IV” by Led Zeppelin, released in 1971, includes famous tracks like “Stairway to Heaven,” blending elements of folk, blues, rock, and heavy metal.
- Taylor Swift’s “1989,” released in 2014, marked her shift from country to pop music.
- “A Night at the Opera” by Queen, released in 1975, is a rock album that includes the hit “Bohemian Rhapsody.”

| AlbumName | Artist | Genre | ReleaseYear |
|---------------------------|-----------------|--------------------------------|-------------|
| Thriller | Michael Jackson | Pop, Post-disco, Rock | 1982 |
| Back in Black | AC/DC | Hard Rock | 1980 |
| Like a Virgin | Madonna | Pop, Dance | 1984 |
| The Dark Side of the Moon | Pink Floyd | Progressive Rock | 1973 |
| Led Zeppelin IV | Led Zeppelin | Folk, Blues, Rock, Heavy Metal | 1971 |
| 1989 | Taylor Swift | Pop | 2014 |
| A Night at the Opera | Queen | Rock | 1975 |

Step1: 1NF

Remove duplicates, ensure each column has one value(genre doesn't). The **Genre** column in our initial flat file contains multiple genres for some albums. We'll solve this by creating separate rows for each genre associated with an album.

| AlbumName | Artist | Genre | ReleaseYear |
|---------------|-----------------|------------|-------------|
| Thriller | Michael Jackson | Pop | 1982 |
| Thriller | Michael Jackson | Post-disco | 1982 |
| Thriller | Michael Jackson | Rock | 1982 |
| Back in Black | AC/DC | Hard Rock | 1980 |
| Like a Virgin | Madonna | Pop | 1984 |

| AlbumName | Artist | Genre | ReleaseYear |
|---------------------------|--------------|------------------|-------------|
| Like a Virgin | Madonna | Dance | 1984 |
| The Dark Side of the Moon | Pink Floyd | Progressive Rock | 1973 |
| Led Zeppelin IV | Led Zeppelin | Folk | 1971 |
| Led Zeppelin IV | Led Zeppelin | Blues | 1971 |
| Led Zeppelin IV | Led Zeppelin | Rock | 1971 |
| Led Zeppelin IV | Led Zeppelin | Heavy Metal | 1971 |
| 1989 | Taylor Swift | Pop | 2014 |
| A Night at the Opera | Queen | Rock | 1975 |

Step2: 2NF

Check for functional dependencies. And create tables

We don't have any functional dependencies. So each column will have their own table.

Artists table:

| ArtistID | Artist |
|----------|-----------------|
| 1 | Michael Jackson |
| 2 | AC/DC |
| 3 | Madonna |
| 4 | Pink Floyd |
| 5 | Led Zeppelin |
| 6 | Taylor Swift |
| 7 | Queen |

Genres table:

| GenreID | Genre |
|---------|------------------|
| 1 | Pop |
| 2 | Post-disco |
| 3 | Rock |
| 4 | Hard Rock |
| 5 | Dance |
| 6 | Progressive Rock |
| 7 | Folk |
| 8 | Blues |
| 9 | Heavy Metal |

Albums table:

| AlbumID | AlbumName |
|---------|---------------------------|
| 1 | Thriller |
| 2 | Back in Black |
| 3 | Like a Virgin |
| 4 | The Dark Side of the Moon |

| AlbumID | AlbumName |
|---------|----------------------|
| 5 | Led Zeppelin IV |
| 6 | 1989 |
| 7 | A Night at the Opera |

We can leave it here. We dont need to add artists, genres, and realease years to Albums table.