SQL - Fusionar conjuntos de datos

en R y SAS

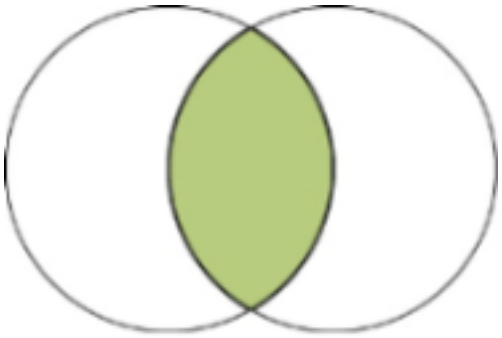
César Mignoni

Introducción

Las formas, que se ha vistos, de ralizar consultas con SQL, es utilizando una sola tabla. Sin embargo, a menudo es necesario obtener datos de tablas independientes. Cuando especifican varias *tablas* de consulta en la cláusula FROM, SQL las procesa para formar un conjunto de datos. El *tablas* resultante contiene datos de cada *tablas* fuente. Estas consultas se conocen como *combinaciones*.

Conceptualmente, cuando se especifican dos conjuntos de datos, SQL hace coincidir cada fila del primero con todas las filas del segundo para producir un conjunto de datos interno o intermedio conocido como el *producto cartesiano*. El producto cartesiano de grandes conjuntos de datos puede ser enorme, pero normalmente se obtienen subconjuntos de datos declarando el tipo de combinación. Hay dos tipos de combinaciones:

- **Las combinaciones internas** devuelven una tabla de resultados para todas las filas de una tabla que tienen una o más filas coincidentes en la otra tabla o tablas.

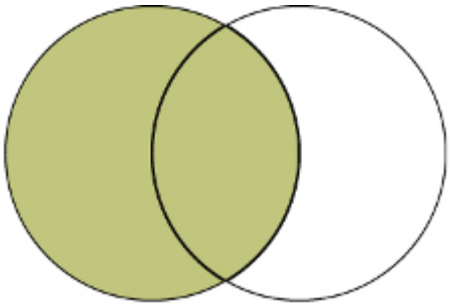


- **Las combinaciones externas** son combinaciones internas que se aumentan con filas que no coinciden con ninguna fila de la otra tabla en la combinación. Hay tres tipos de combinaciones externas: izquierda, derecha y completa.

La tarea central es unir tablas para obtener un detalle sobre la consulta de tablas individuales.

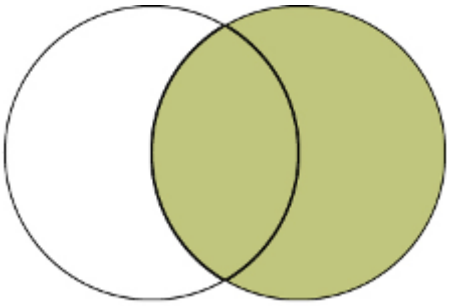
Unión izquierda

Es un tipo de *unión exterior* en la que la tabla de resultados incluye todas las observaciones de la tabla izquierda, ya sea o no se encuentra una coincidencia para ellos en ninguna de las tablas especificadas a la derecha. Una combinación izquierda entre dos tablas puede ser representada gráficamente como se muestra en el siguiente diagrama de Venn.



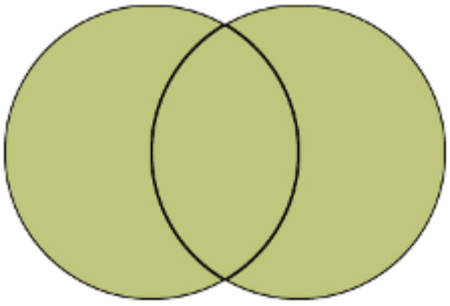
Unión derecha

Es idéntica a la *unión izquierda*, excepto que la tabla de resultados incluye todas las observaciones de la tabla derecha, si se encuentra o no una coincidencia para ellos en cualquiera de las tablas especificadas a la izquierda.



Unión completa

Cuando una unión se especifica como una combinación completa, la tabla de resultados incluye todas las observaciones del producto cartesiano de dos tablas para las cuales la expresión SQL es verdadera, más las filas de cada tabla que no coinciden con ninguna fila en la otra tabla. La representación visual de la unión externa completa se muestra en el siguiente diagrama de Venn.



Fusión de conjuntos de datos (interna y externa) en R

Se puede combinar dos conjuntos de datos en R usando la función *Merge()*. Los conjuntos de datos deben tener los mismos nombres de columna en los que se produce la fusión. La función *Merge()* en R es similar a la operación de unión de tablas en base de datos con SQL. Los diferentes argumentos para *Merge()* permiten realizar combinaciones naturales, así como

combinaciones externas izquierda, derecha y completa.

Los argumentos de la función *merge()* son:

- **"x"** : conjunto de datos 1.
- **"y"** : conjunto de datos 2.
- **"by"** , **"x"** , **"by.y"** : los nombres de las columnas que son comunes a **"x"** e **"y"** . El valor predeterminado es usar las columnas con nombres comunes entre los dos conjuntos de datos.
- **"all,all.x,all.y"** : valores lógicos que especifican el tipo de fusión. El valor predeterminado es "all=FALSE" (lo que significa que solo se devuelven las filas coincidentes).

Cómo definir los argmentos para los diferentes tipos de fusión:

- ***Unión natural (interna)*** : para mantener solo las filas que coinciden con los conjuntos de datos, especifique el argumento *"all=FALSE"*.
- ***Unión externa completa***: para mantener todas las filas de ambos conjuntos de datos, especifique *"all=TRUE"* .
- ***Unión externa izquierda***: para incluir todas las filas de su conjunto de datos **"x"** y solo aquellas de **"y"** que coincidan, especifique *"all.x=TRUE"*.
- ***Unión externa derecha***: para incluir todas las filas de su conjunto de datos **"y"** y solo las de **"x"** que coinciden, especifique *"all.y=TRUE"* .

Ejemplos prácticos.

Considerando dos conjuntos de datos

Data.frame_1

```
df1 = data.frame(CustomerId = c(1:6), Product = c(rep("Horno", 3), rep("Televisión", 3))
df1
```

```
##   CustomerId    Product
## 1          1      Horno
## 2          2      Horno
## 3          3      Horno
## 4          4 Televisión
## 5          5 Televisión
## 6          6 Televisión
```

Data.frame_2

```
df2 = data.frame(CustomerId = c(2, 4, 6), State = c(rep("California", 2), rep("Texas", 1
df2
```

```
##   CustomerId      State
## 1          2 California
## 2          4 California
## 3          6      Texas
```

Unión interna: devuelve solo las filas en las que la tabla izquierda tiene claves coincidentes en la tabla derecha.

```
df <- merge(x = df1, y = df2, by = "CustomerId")
```

El conjunto de datos "df" resultante será:

```
##   CustomerId    Product      State
## 1          2      Horno California
## 2          4 Televisión California
## 3          6 Televisión      Texas
```

Unión externa: devuelve todas las filas de ambas tablas, registros de unión de la izquierda que tienen claves coincidentes en la tabla derecha.

```
df <- merge(x = df1, y = df2, by = "CustomerId", all = TRUE)
```

El conjunto de datos" df" resultante será:

```
##   CustomerId    Product      State
## 1          1      Horno       <NA>
## 2          2      Horno California
## 3          3      Horno       <NA>
## 4          4 Televisión California
## 5          5 Televisión       <NA>
## 6          6 Televisión      Texas
```

Unión externa izquierda: devuelve todas las filas de la tabla izquierda y cualquier fila con teclas coincidentes de la tabla derecha.

```
df <- merge(x = df1, y = df2, by = "CustomerId", all.x = TRUE)
```

El conjunto de datos" df" resultante será:

```
##   CustomerId    Product      State
## 1          1      Horno       <NA>
## 2          2      Horno California
## 3          3      Horno       <NA>
## 4          4 Televisión California
## 5          5 Televisión       <NA>
## 6          6 Televisión      Texas
```

Unión externa derecha: devuelve todas las filas de la tabla derecha y cualquier fila con teclas coincidentes de la tabla izquierda.

```
df <- merge(x = df1, y = df2, by = "CustomerId", all.y = TRUE)
```

El conjunto de datos" df" resultante será:

```
##   CustomerId    Product      State
## 1          2      Horno California
## 2          4 Televisión California
## 3          6 Televisión      Texas
```

Combinación cruzada: una combinación cruzada (también conocida como combinación cartesiana) da como resultado que cada fila de una tabla se una a cada fila de otra tabla.

```
df<-merge(x = df1, y = df2, by = NULL)
```

El conjunto de datos" df" resultante será:

```
##    CustomerId.x    Product CustomerId.y      State
## 1             1      Horno            2 California
## 2             2      Horno            2 California
## 3             3      Horno            2 California
## 4             4 Televisión            2 California
## 5             5 Televisión            2 California
## 6             6 Televisión            2 California
## 7             1      Horno            4 California
## 8             2      Horno            4 California
## 9             3      Horno            4 California
## 10            4 Televisión            4 California
## 11            5 Televisión            4 California
## 12            6 Televisión            4 California
## 13            1      Horno            6      Texas
## 14            2      Horno            6      Texas
## 15            3      Horno            6      Texas
## 16            4 Televisión            6      Texas
## 17            5 Televisión            6      Texas
## 18            6 Televisión            6      Texas
```

Fusión de conjuntos de datos (interna y externa) en SAS

PROC SQL implementa el lenguaje de consulta estándar y permite al usuario la unión de *dataset* mediante consultas de combinación.

Como se ha descripto en el material "*Lenguaje de Consulta Estructurado en SAS y R*", en **PROC SQL** la cláusula *FROM* se utiliza en una expresión de consulta para especificar el/los conjunto(s) de datos fuente, y que se combinan para producir el resultado de la unión.

Además de los diversos tipos de combinaciones (internas y externas) que se describen, los ejemplos que se incluyen muestran la igualdad entre los valores de columna provenientes de las tablas que se están uniendo; comparación entre valores calculados; etc. La cláusula *WHERE* o la cláusula *ON* contienen las condiciones bajo las cuales algunas filas son guardadas o eliminadas en la tabla de resultados. *WHERE* se usa para seleccionar filas de uniones internas. *ON* se utiliza para seleccionar filas de uniones internas o externas.

Ejemplos

Unión interna

```
proc sql;
   title 'Oil Production/Reserves of Countries';
   select p.country, barrelsperday 'Production',
          barrels 'Reserves'
      from oilprod p, oilrsrvs r
      where p.country = r.country
      order by barrelsperday desc;
quit;
```

Unión externa izquierda

```
proc sql;
   title 'Coordinates of Capital Cities';
```

```
   select Capital format=$20.,
          Name 'Country' format=$20.,
          Latitude, Longitude
      from countries a left join
           worldcitycoords b
           on a.Capital = b.City and
              a.Name = b.Country
      order by Capital;
quit;
```

Unión externa derecha

```
proc sql;
   title 'Populations of Capitals Only';
   select City format=$20.,
          Country 'Country' format=$20.,
          Population
      from countries right join
           worldcitycoords
           on Capital = City and
              Name = Country
      order by City;
quit;
```

Unión externa completa

```
proc sql;
   title 'Populations/Coordinates of World Cities';
   select City '#City#(WORLDCITYCOORDS)' format=$20.,
          Capital '#Capital#(COUNTRIES)' format=$20.,
          Population, Latitude, Longitude
      from countries full join
           worldcitycoords
           on Capital = City and
              Name = Country;
quit;
```

Borrar los títulos

```
title;
```