10

Para saber mais

Quando falamos sobre cruzamentos de bases, reforçamos que era importante garantir que as chaves de cruzamento não tenham duplicações. Mas o que acontece se elas tiverem?

Vamos aqui usar duas bases de dados hipotéticas, duas agendas de contatos, uma com **Nome** e **Telefone**, outra com **Nome** e **Email**, assim, a chave de cruzamento delas é o nome. Nelas temos duplicações, pois "Beltrano" tem dois números de telefone, "Fulano" tem dois endereços de email e "Sicrano" tem dois telefones e dois emails. Essas duas bases estão exemplificadas abaixo:

Nome	Telefone
Beltrano	Número 1
Beltrano	Número 2
Fulano	Número 1
Sicrano	Número 1
Sicrano	Número 2

Nome	Email
Beltrano	Endereço 1
Fulano	Endereço 1
Fulano	Endereço 2
Sicrano	Endereço 1
Sicrano	Endereço 2

Como a base resultante do cruzamento dessas duas depende se iremos fazer este cruzamento usando um *data step* com *merge* ou um *proc SQL* com *join*.

SQL

O SQL faz o que normalmente é chamado de *"união cruzada"* ou *"produto cartesiano"*, no sentido que a base resultante possui uma linha para cada vez que a chave de uma base encontrar uma correspondência (*"match"*, em inglês) na outra base.

Por exemplo, tomemos a primeira linha da primeira base, onde a chave é "Beltrano" e temos o primeiro número de telefone dele; varrendo toda a segunda base, encontramos uma correspondência para a chave "Beltrano", com o primeiro endereço de email dele. Na segunda linha da base a chave é novamente "Beltrano", e o processo de varrer toda a segunda base é feito novamente, onde mais uma vez é encontrada uma correspondência.

Agora, no caso do "Sicrano" a primeira linha dele na primeira base (com o primeiro número de telefone) encontra duas correspondências na segunda (pois ele tem dois emails), logo isso gera duas linhas na base. A segunda vez que "Sicrano" aparece também encontra duas correspondências, o que gera mais duas linhas. Assim, na base final temos "Sicrano" aparecendo um total de 4 vezes, com o primeiro telefone correspondido com os dois emails e o segundo telefone também correspondido com os 2 emails. A base resultante do cruzamento por SQL segue abaixo:

Nome	Telefone	Email
Beltrano	Número 1	Endereço 1
Beltrano	Número 2	Endereço 1
Fulano	Número 1	Endereço 1
Fulano	Número 1	Endereço 2
Sicrano	Número 1	Endereço 1
Sicrano	Número 1	Endereço 2
Sicrano	Número 2	Endereço 1
Sicrano	Número 2	Endereço 2

Se "Sicrano" tivesse 3 números de telefone, ele iria aparecer 6 vezes na base final, com cada um dos 3 telefones pareados com cada um dos 2 emails, e assim por diante.

MERGE

O merge segue um processo diferente. Ele busca garantir que todas as informações de ambas as bases estejam presentes na base final e não que todas as correspondências estejam representadas. Isso faz com que, na base gerada pelo cruzamento, a chave de cruzamento tenha o número máximo de duplicações que essa chave tem em cada uma das bases participantes.

Ou seja, no nosso exemplo, “Beltrano” aparece duas vezes na primeira base e uma vez na segunda, logo na base resultante teremos “Beltrano” duas vezes (o máximo de vezes que ele aparece em alguma das bases), e assim por diante, até chegarmos que na base final teremos 6 observações, com “Beltrano”, “Fulano” e “Sicrano”, cada um, aparecendo duas vezes.

Como “Beltrano” possui dois números de telefone, eles aparecem em cada uma das duas linhas, e o único email que ele tem é repetido em ambas as linhas. No caso do “Sicrano”, que também aparece em duas linhas, cada linha tem um telefone e um email diferente. O resultado do cruzamento por *merge* segue abaixo:

Nome	Telefone	Email
Beltrano	Número 1	Endereço 1
Beltrano	Número 2	Endereço 1
Fulano	Número 1	Endereço 1
Fulano	Número 1	Endereço 2
Sicrano	Número 1	Endereço 1
Sicrano	Número 2	Endereço 2

Se “Sicrano” tivesse 3 telefones, ele iria aparecer 3 vezes na base final, cada linha com um telefone diferente, na primeira linha onde ele aparece estaria o primeiro email, na segunda linha estaria o segundo email e na terceira, como ele não tem mais emails, o segundo email apareceria mais uma vez.

Conclusão

Os resultados de cruzamentos com duplicações podem ter resultados pouco intuitivos e complicados de se utilizar. No merge, apesar de ter todas as informações, não há controle completo de como elas serão pareadas (por exemplo, não tenho nenhuma linha com o primeiro telefone e o segundo email do “Sicrano”), podendo ser difícil de encontrar a correspondência específica que se precisa. Já no SQL teremos todas as correspondências possíveis, o que também pode ser um problema pois a base resultante pode ter muitas linhas adicionais; mesmo neste pequeno exemplo a base resultante já tem quase o dobro do tamanho de cada uma das bases cruzadas, e isso pode ficar impraticável se as bases a serem cruzadas são muito grandes.

Isso é uma das características do cruzamento de bases com chaves duplicadas, o resultado normalmente é maior que as bases iniciais. Assim, se após algum cruzamento você se deparar com uma base que é muito grande, a provável causa são duplicações na chave de cruzamento.

Cruzamentos são algo muito importante e útil na análise de dados, então é bom sempre tomar cuidado ao usar este recurso.