

Particionamento: Repartition vs Coalesce

Operação	Repartition	Coalesce
Mecânica	Full Shuffle	Redução local
Uso Ideal	Aumentar paralelismo	Reduzir partições
Performance	Mais lenta	Mais rápida
Impacto	Tamanhos uniformes	Tamanhos desiguais

Repartition vs Coalesce

Dataset de exemplo (dados distribuídos)

Dataset completo

id	valor
1	10
2	20
3	30
4	40
5	50
6	60
7	70
8	80

Distribuído em 4 partições (original)

Partição 1

id	valor
1	10
2	20

Partição 2

id	valor
3	30
4	40

Partição 3

id	valor
5	50
6	60

Partição 4

id	valor
7	70
8	80

1 REPARTITION (com shuffle)

```
df.repartition(2)
```

Após repartition(2)

Partição 1 (nova)

id	valor
1	10
3	30
5	50
7	70

Partição 2 (nova)

id	valor
2	20
4	40
6	60
8	80

- ✓ Redistribui os dados entre todas as partições
- ✓ Envolve shuffle (comunicação entre nós)
- ✓ Balanceia os dados

O que aconteceu?

- Todos os dados foram embaralhados (shuffle)
- Novas partições receberam dados misturados
- Partições balanceadas
- ✓ **Uso típico:** alterar número de partições com balanceamento

Quando usar?

Quando você quer redistribuir os dados para balancear a carga ou aumentar/diminuir paralelismo de forma eficiente.

2 COALESCE (sem shuffle)

```
df.coalesce(2)
```

Após coalesce(2)

Partição 1 (nova)

id	valor
1	10
2	20
3	30
4	40

Partição 2 (nova)

id	valor
5	50
6	60
7	70
8	80

- ✓ Reduz o número de partições sem shuffle
- ✓ Não redistribui os dados
- ✓ Rápido, mas pode gerar desbalanceamento

O que aconteceu?

- As partições foram mescladas (1+2 e 3+4)
- Não houve redistribuição de dados
- Pode ficar desbalanceado se as partições originais não tinham tamanhos parecidos
- ✓ **Uso típico:** reduzir número de partições rapidamente

Quando usar?

Quando você quer reduzir o número de partições rapidamente, especialmente antes de salvar os dados.

Resumo rápido

Aspecto	Repartition	Coalesce
Envolve shuffle?	Sim	Não
Redistribui os dados?	Sim	Não
Balanceamento?	Sim (tende a balancear)	Não (pode desbalancear)
Performance	Mais lento (shuffle)	Mais rápido
Uso típico	Alterar paralelismo com balanceamento	Reduzir partições rapidamente

Dica de ouro

Use `repartition()` quando precisa redistribuir e balancear.

Use `coalesce()` quando só quer reduzir o número de partições de forma rápida.