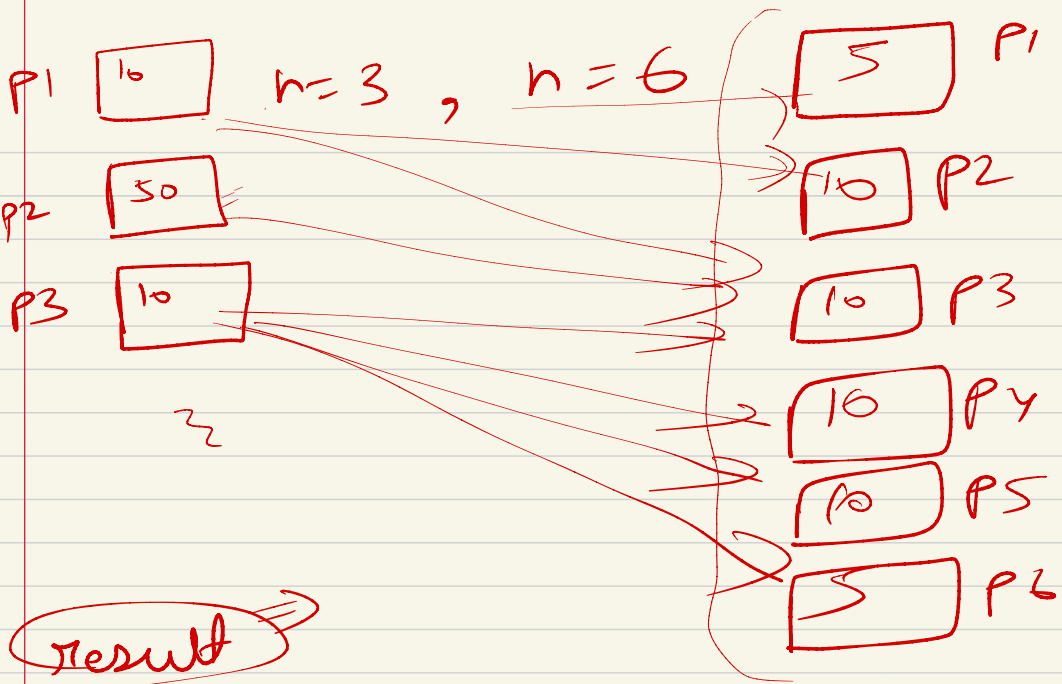


result

result.coalesce(1).write()

↓ (1)

05 x 100

5-

num = 100

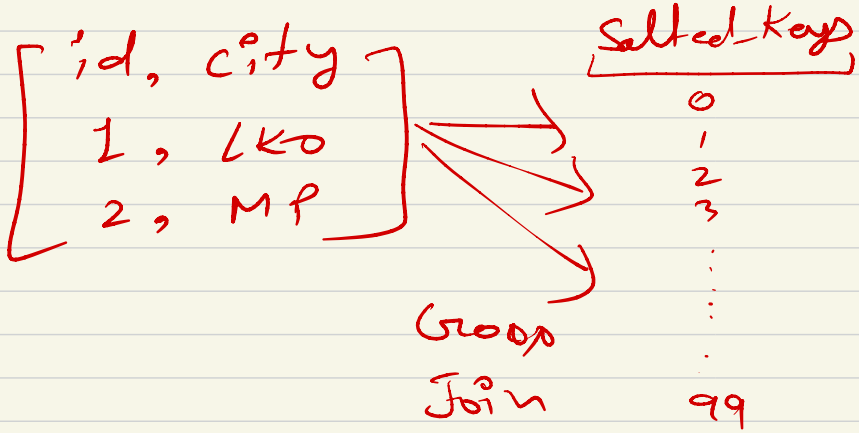
df1 ⇒ (id, name, dept)

df1 ⇒ (id, name, dept, salted_key)

id	name	dept	salted_key
1	a	x	50
2	b	y	10
1	a	x	12

[1, a, x, 18
  2, b, y, 2]

df2 ⇒ (id, city)



df2-exploded =

id	city	salted_key
1	LKO	0
1	LKO	1
1	LKO	...
1	LKO	12
1	LKO	...
1	LKO	18
1	LKO	50
1	LKO	... 99

2, MP, 0
2, MP, 1
2, MP, 2
2, MP, 10
2, MP, 99

df1 =>

id	name	dept	saltedkey
1	X	a	50
2	Y	b	2
1	X	a	12
1	X	a	18
2	Y	b	10

df2-exploded

id	city	saltedk
1	LKO	0
1	LKO	12
1	LKO	18
1	LKO	50
2	MP	2
2	MP	10
2	MP	99

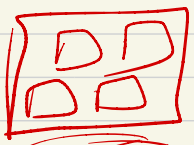
```
from df1
inner join df2-exploded
on df1.id = df2.id and df1.salt
                    = df2.salt
```

--master local[*]

↳ All vcores

n

↳ n vcores



16

spark-submit

--master yarn

--num-executor 10

--executor-memory 4G

63

5

User memory

= usable * (1 - 0.6)

= 3700 * (0.4)

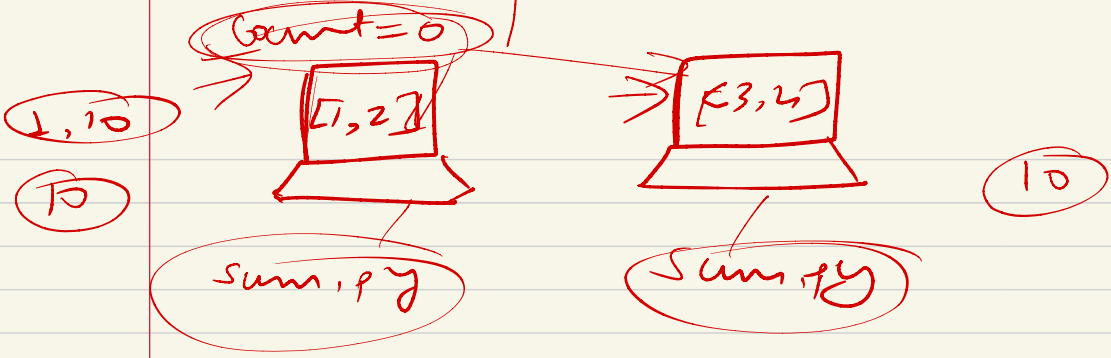
= 370 * 4

= 1480 MB

1268

0 + 1 + 2 = 3

+3 + 4


count=0
[1,2]
[3,4]
sum.py
sum.py