Восстановление распределенной базы данных после аварии

Антон Виноградов

Apache Ignite Committer & PMC Member

github.com/anton-vinogradov/talks

This talk represents my own personal view and opinion.

It does not necessarily reflect the official stance of The Apache Software Foundation/SberTech/any company/any other entity the author might be affiliated with at the moment of presenting or in the past.

Антон Виноградов

Apache Ignite Committer & PMC Member

github.com/anton-vinogradov/talks

Антон Виноградов

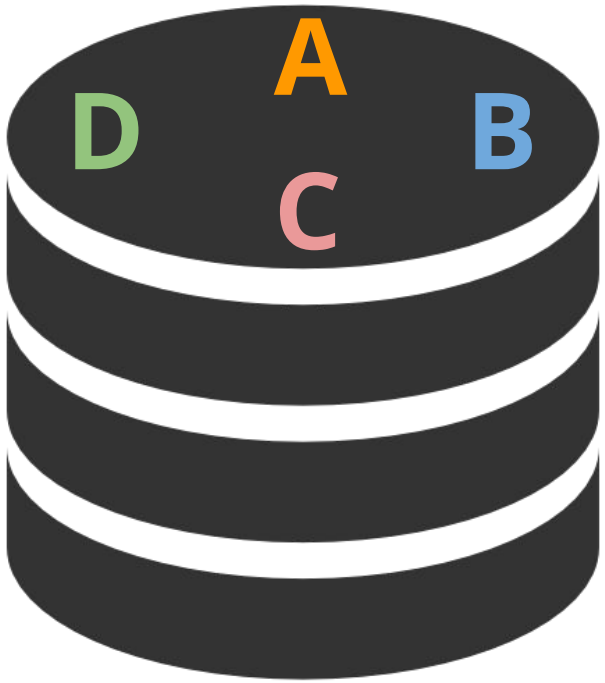
Apache Ignite Committer & PMC Member

github.com/anton-vinogradov/talks

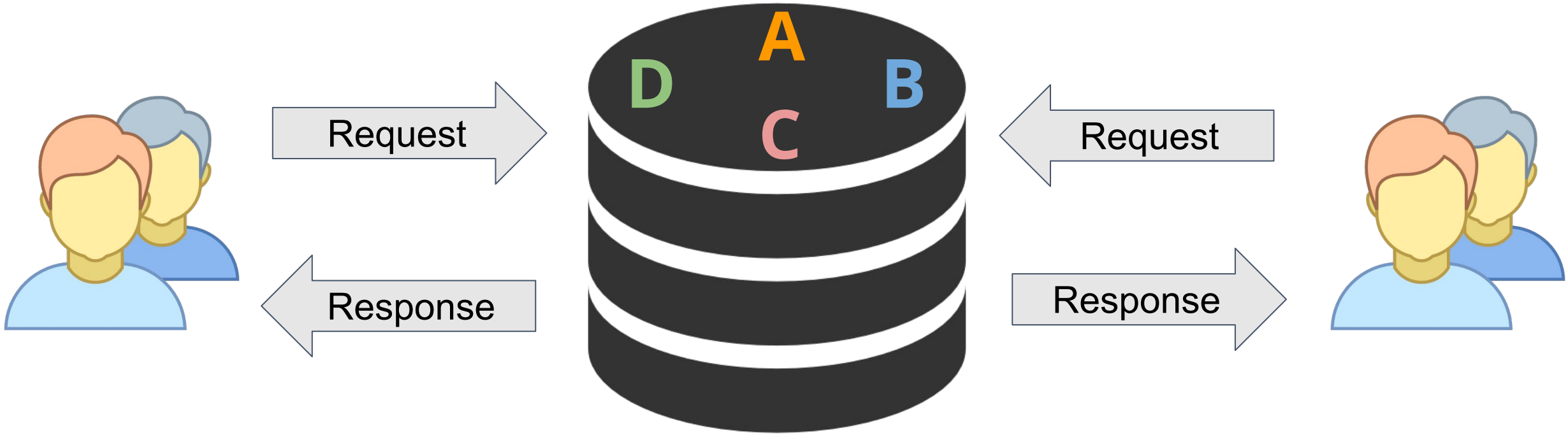

THIS IS FINE.

База данных

База данных



База данных

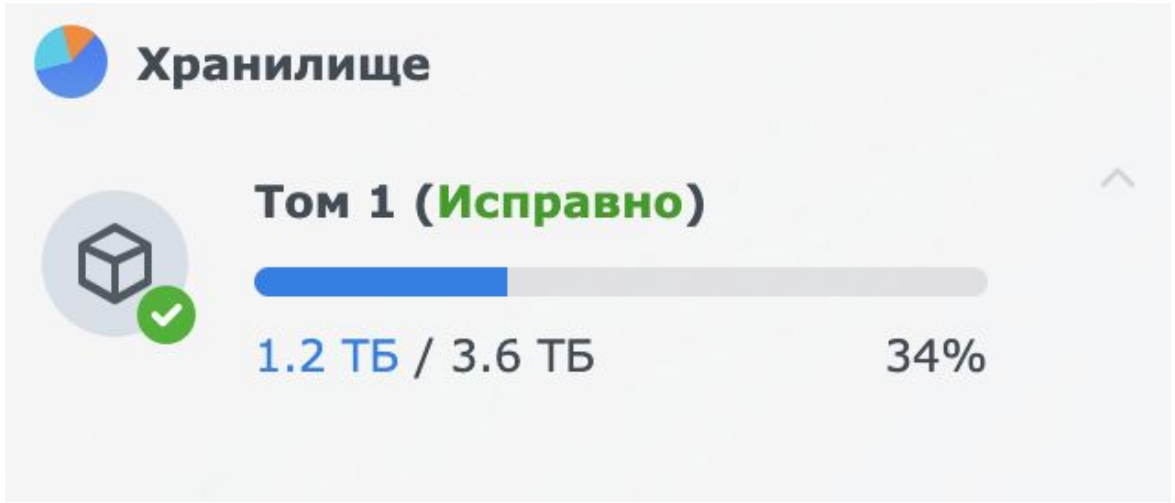


База данных

Не реклама!

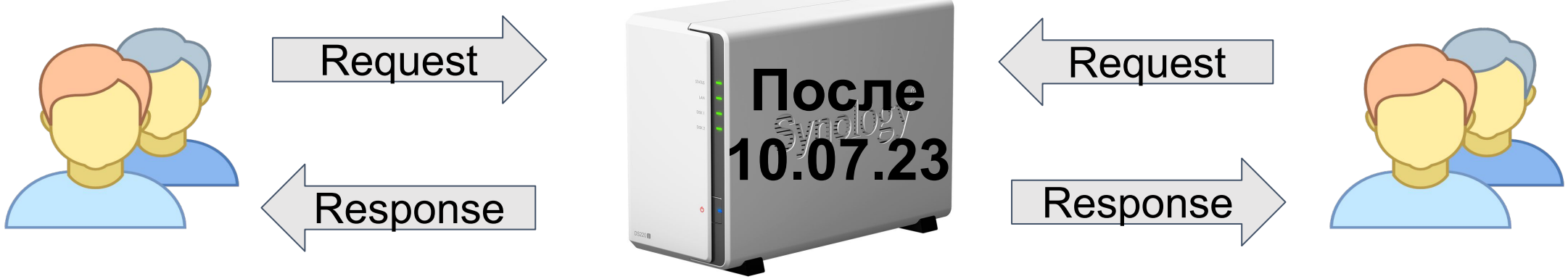
Не реклама!


STATUS
LAN
DISK 1
DISK 2
Synology
DS220

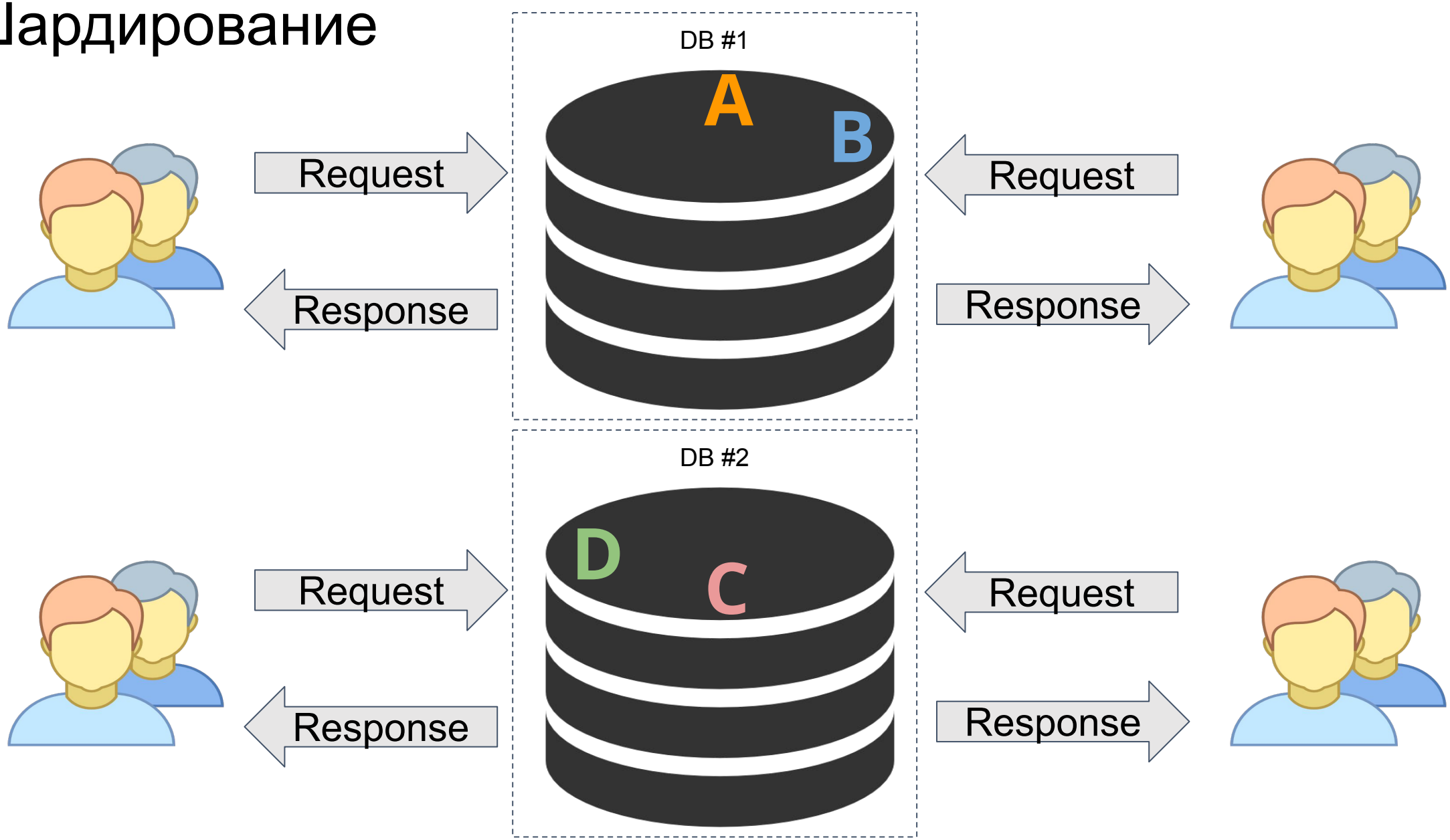

Хранилище
Том 1 (Исправно)
1.2 ТБ / 3.6 ТБ
34%

Шардирование

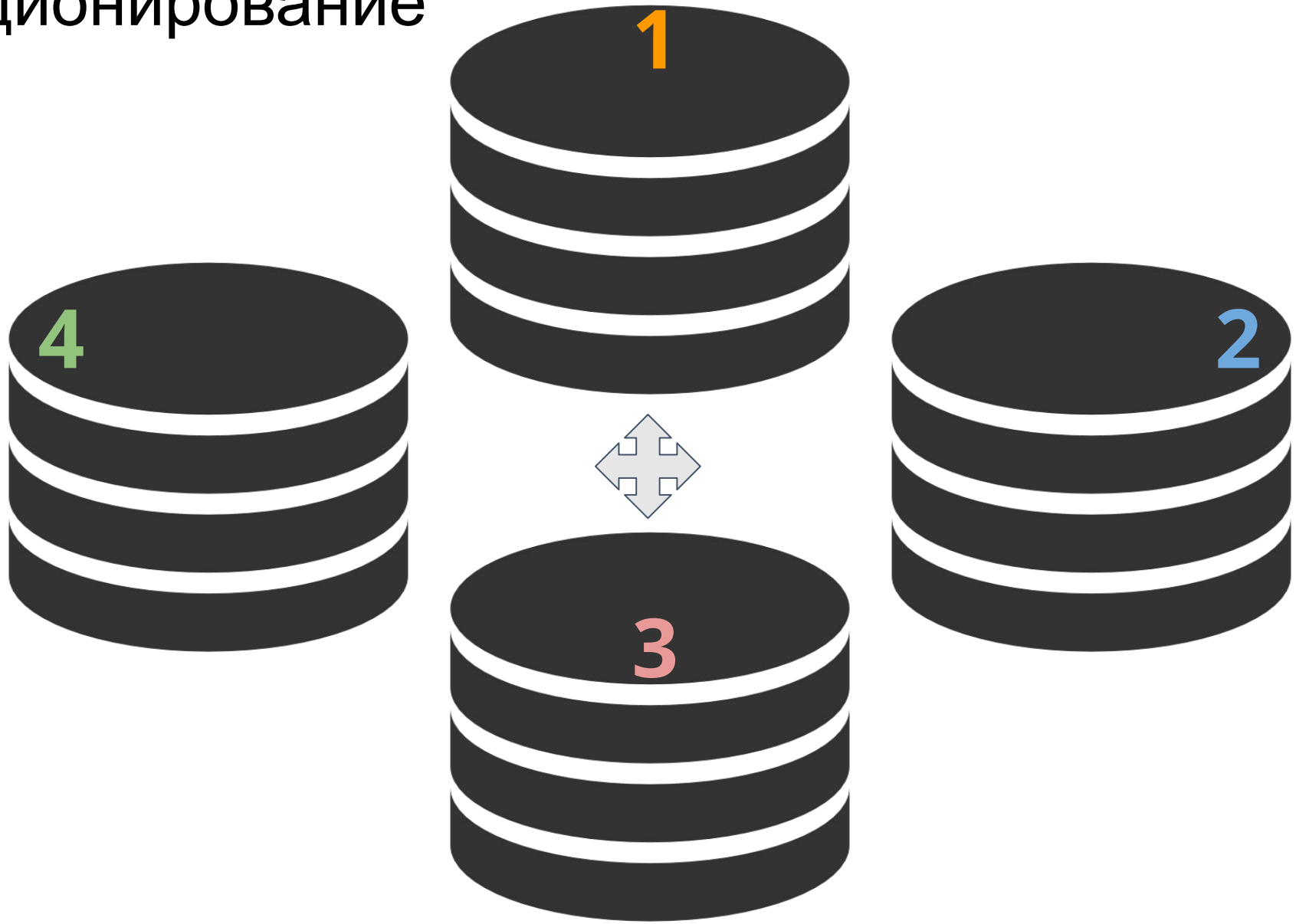
Шардирование



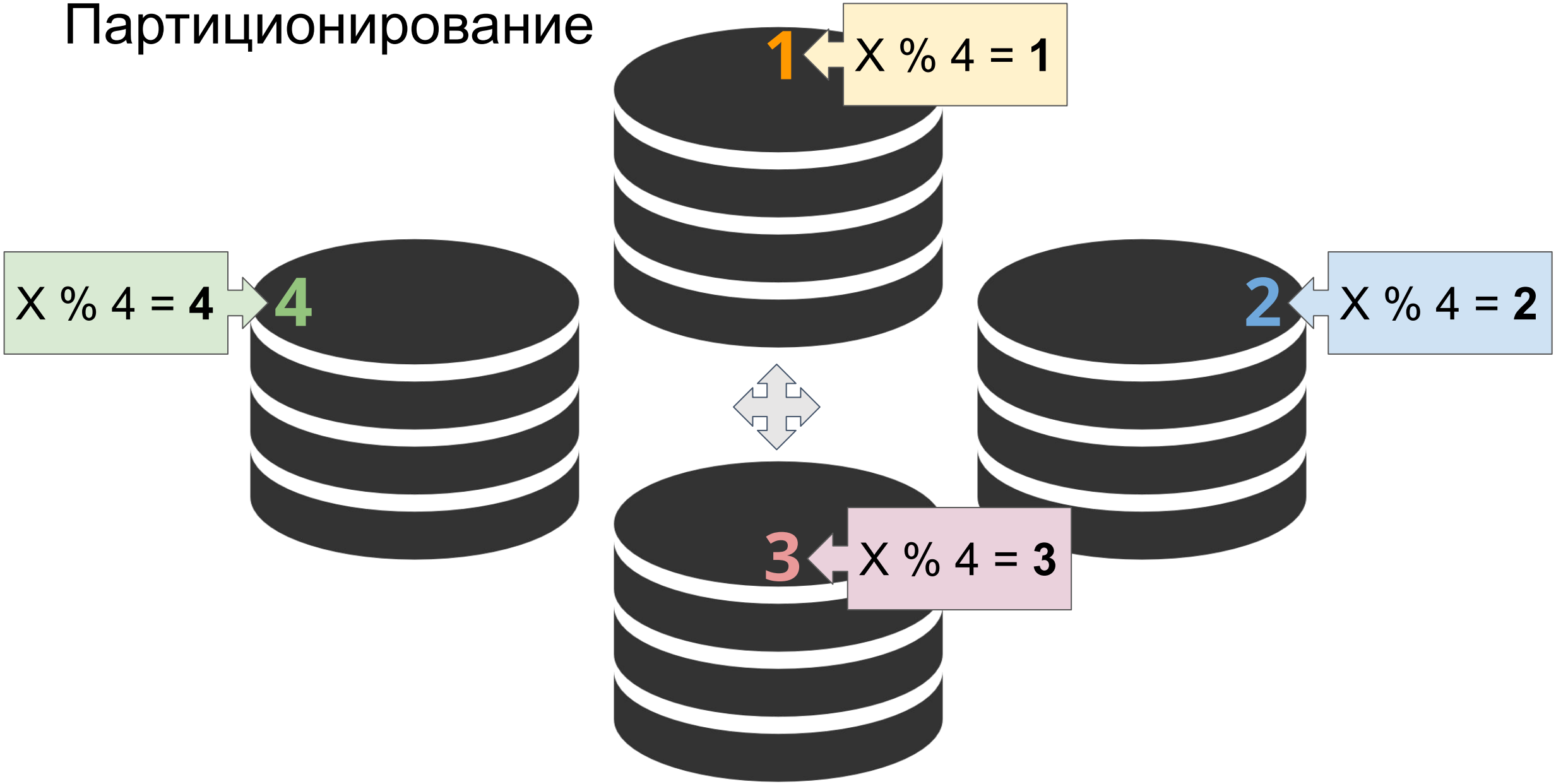
Шардирование



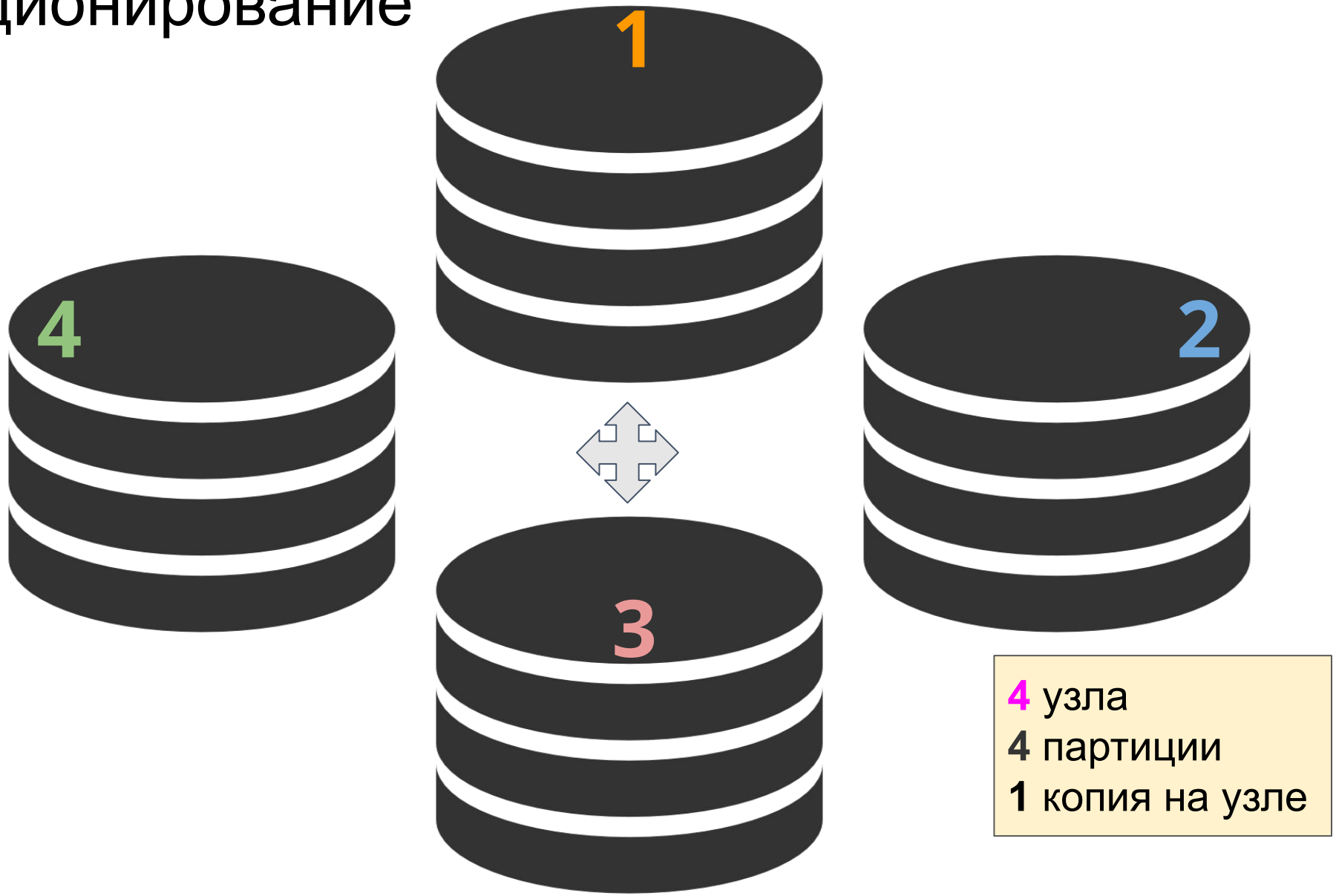
Партиционирование



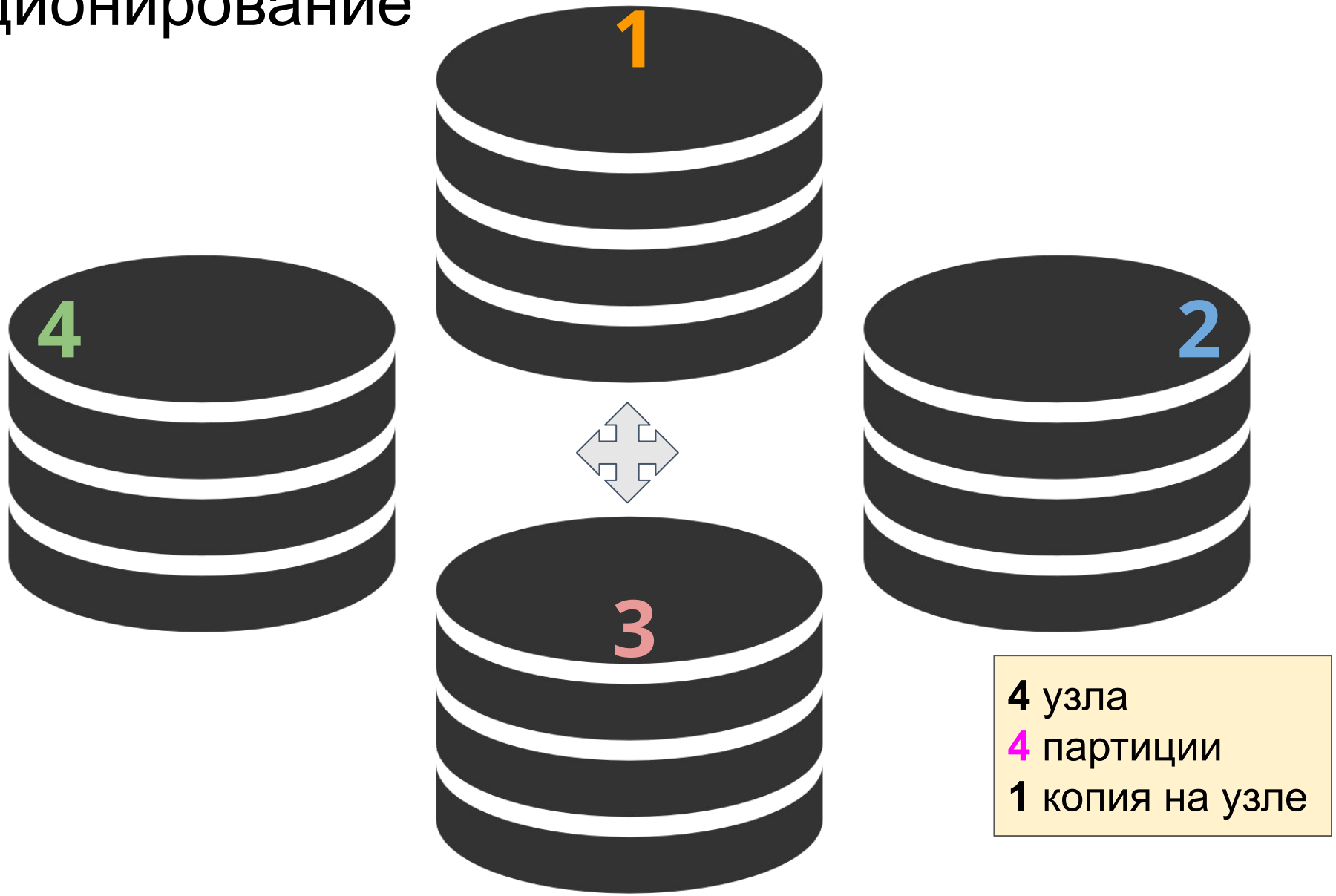
Партиционирование



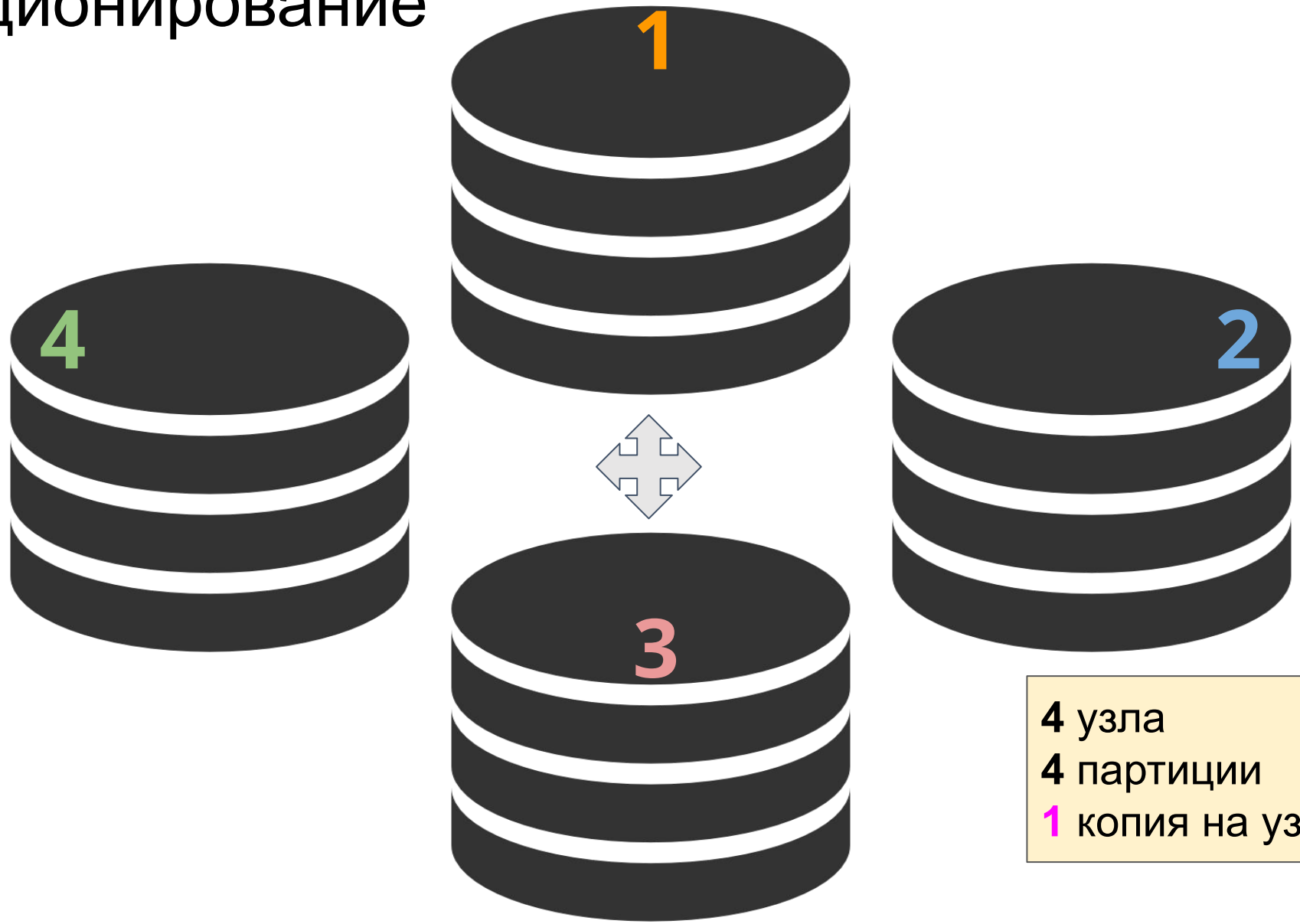
Партиционирование



Партиционирование



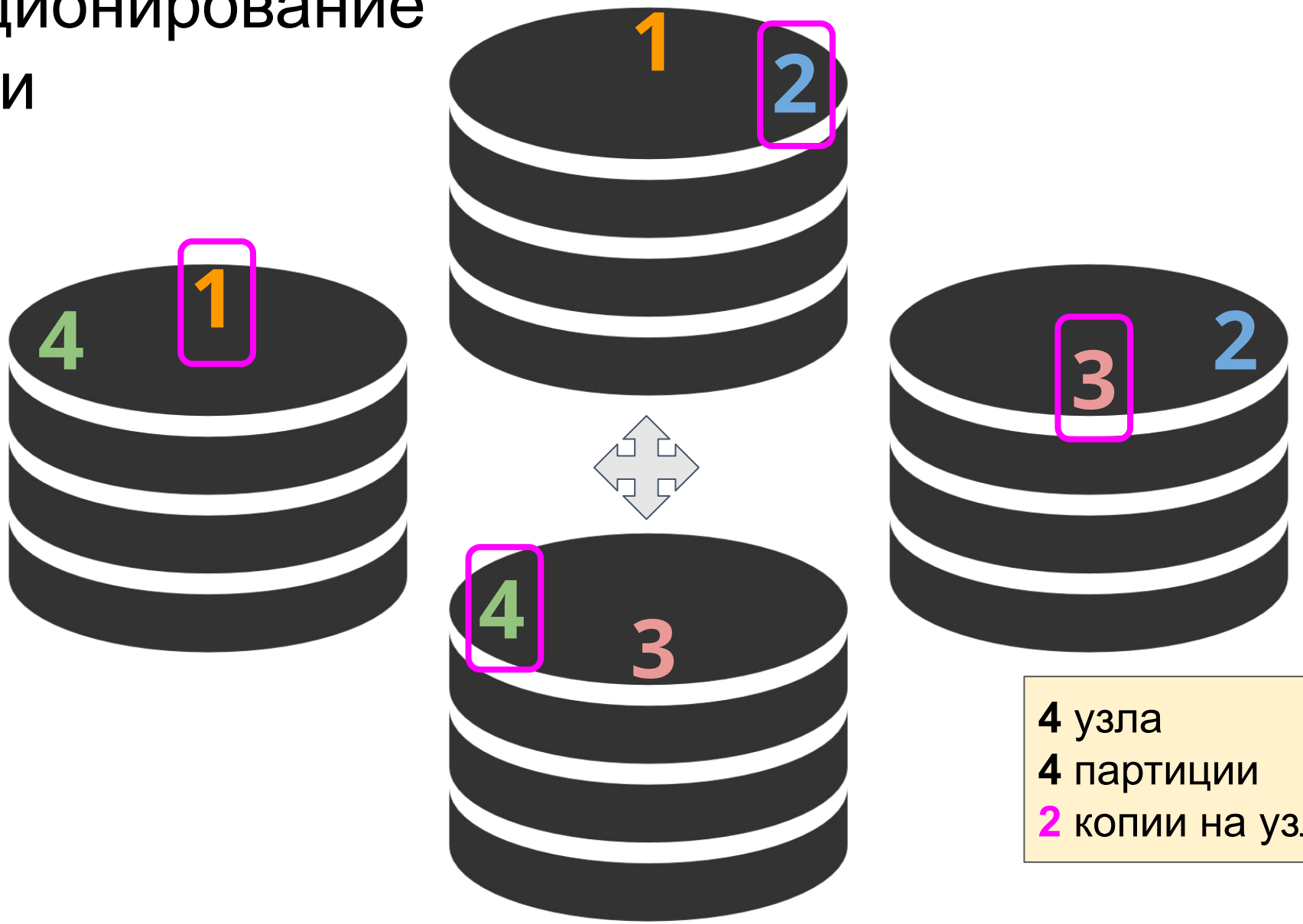
Партиционирование



4 узла
4 партиции
1 копия на узле

1	
2	
3	
4	

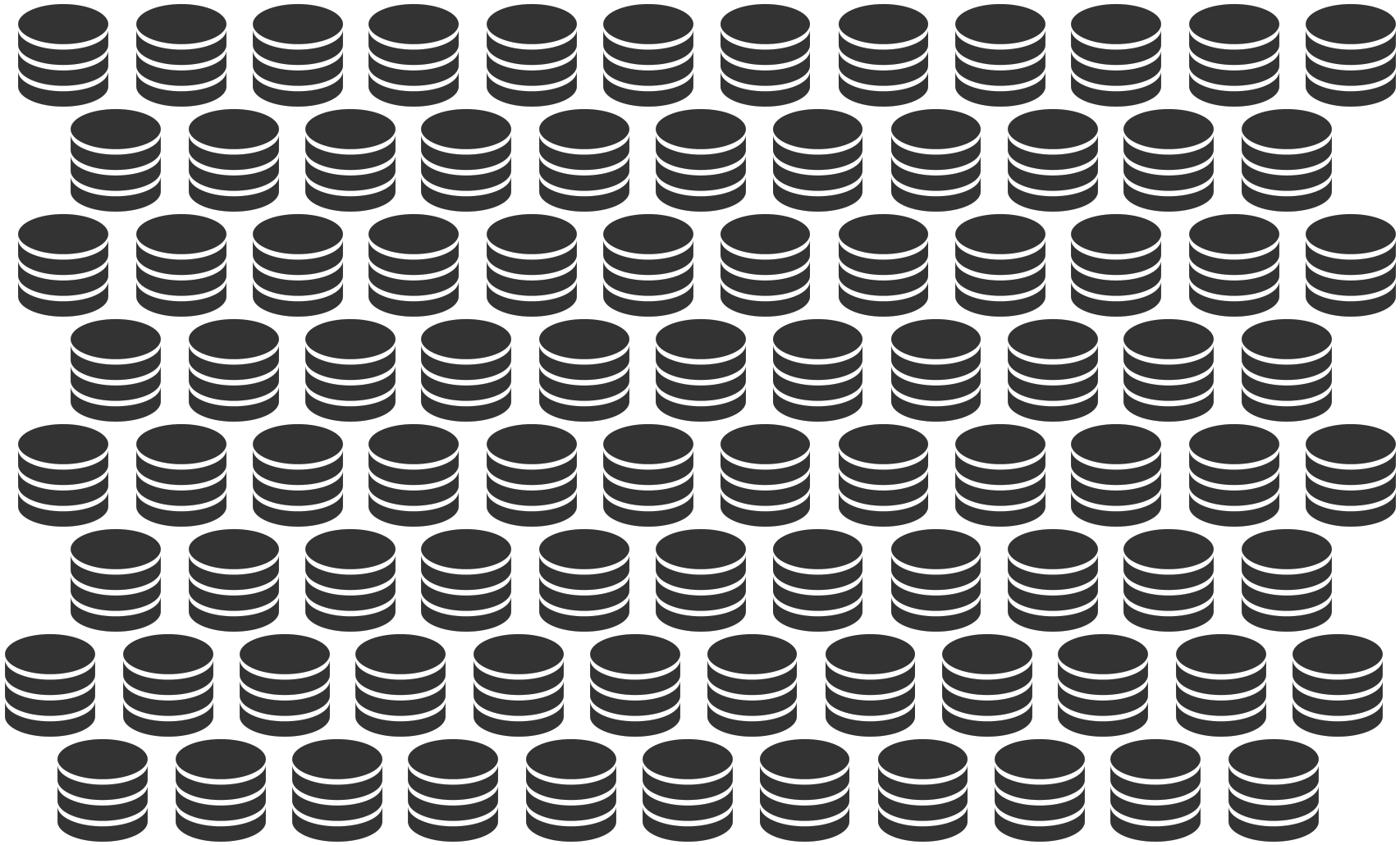
Партиционирование
+ копии



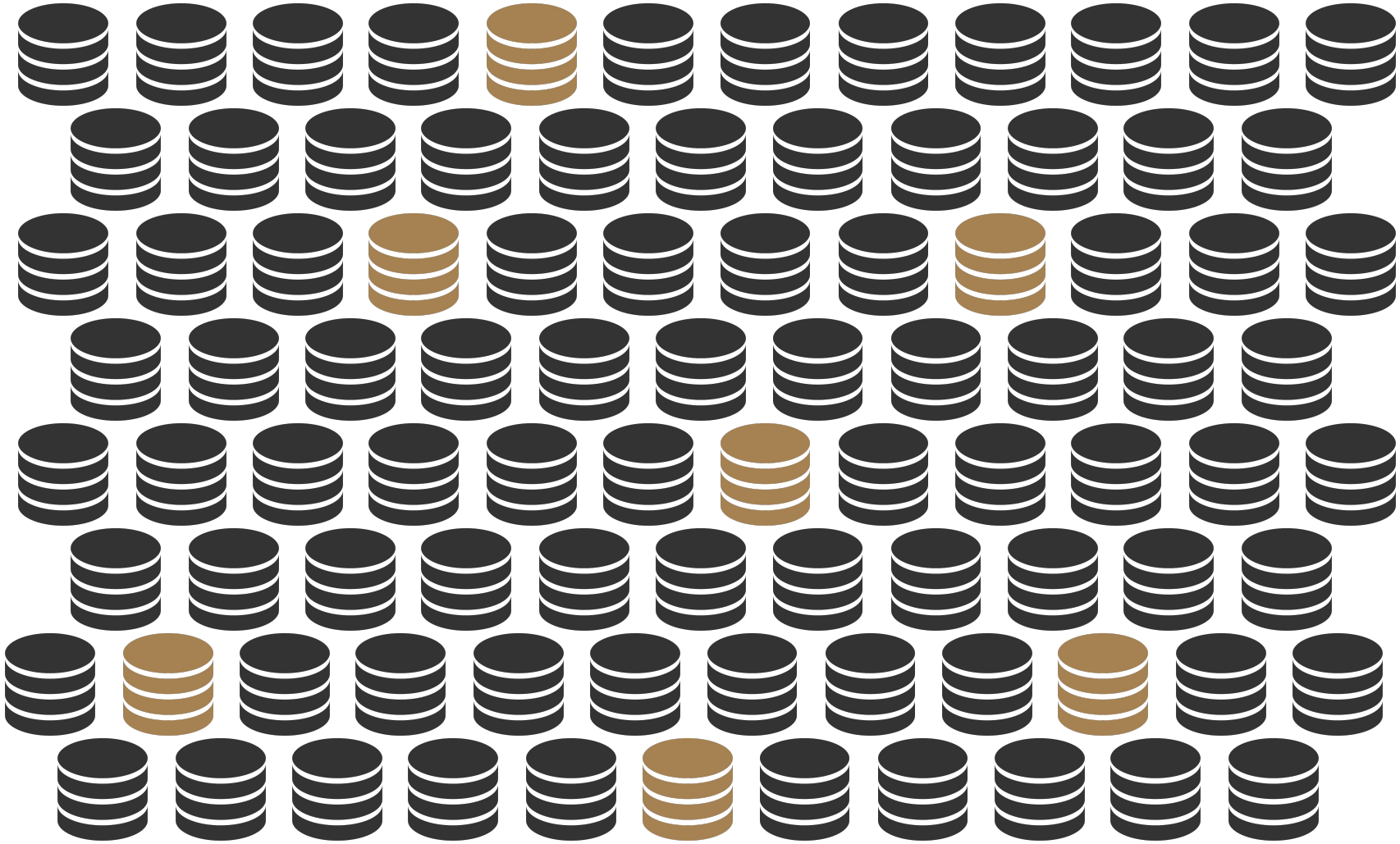
4 узла
4 партиции
2 копии на узле

1	2
2	3
3	4
4	1

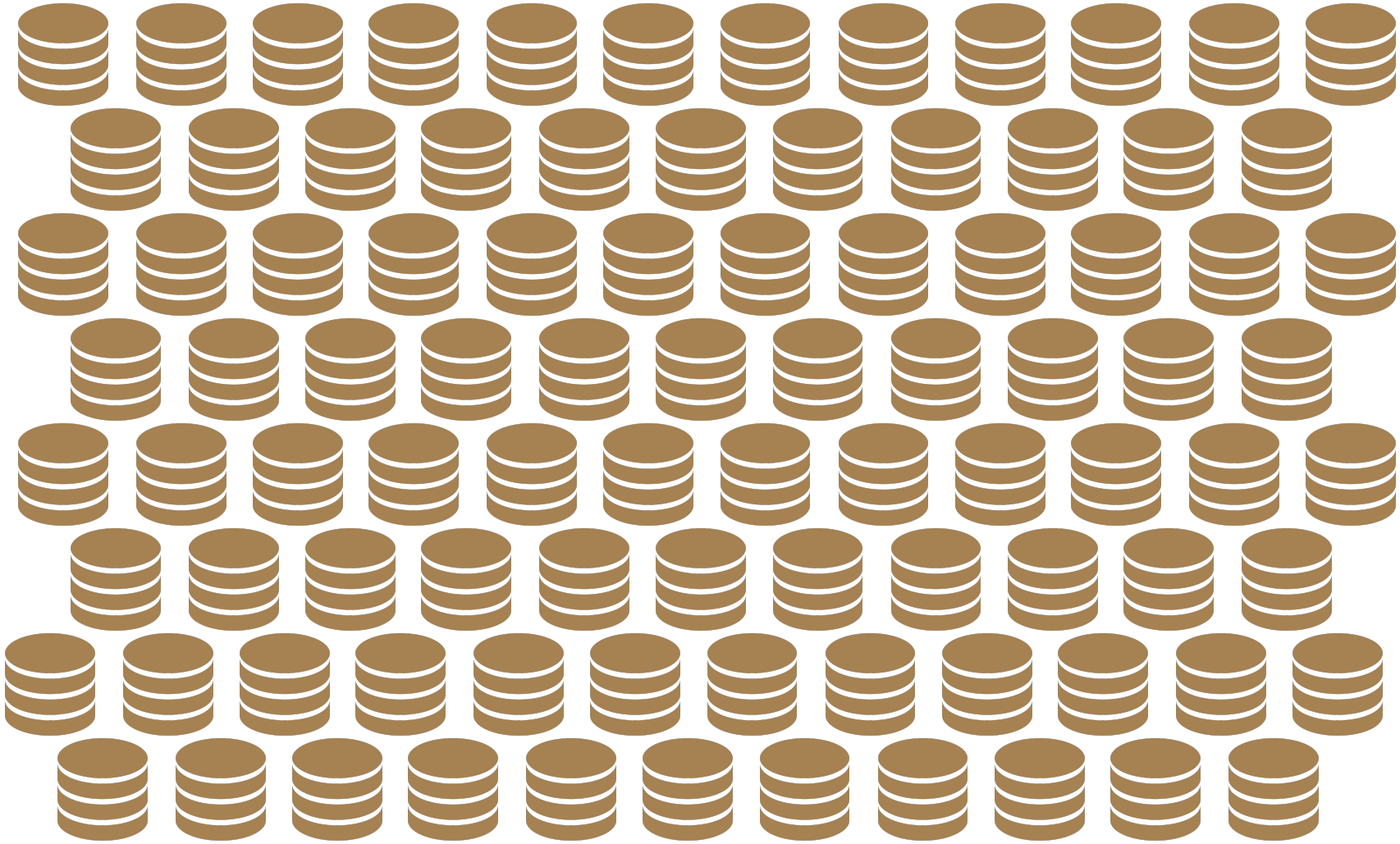
Реальный кластер



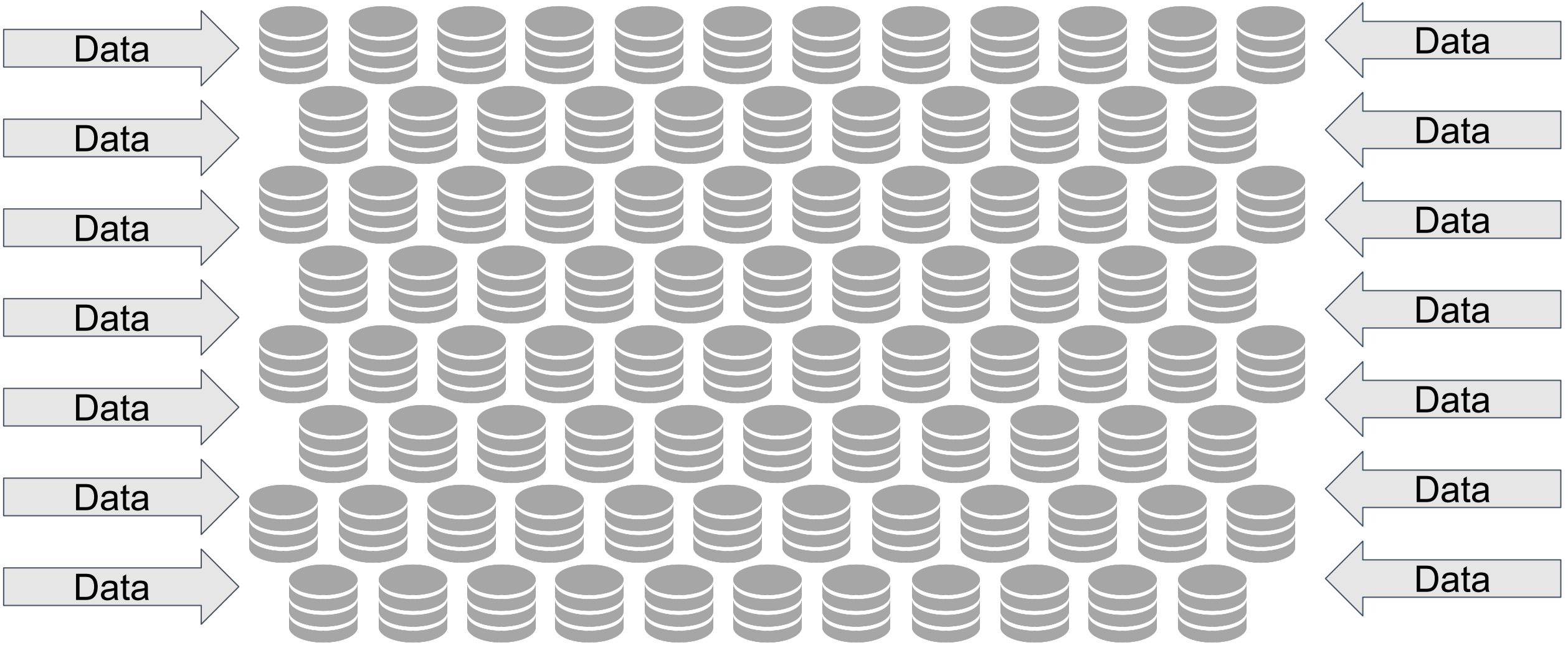
Авария



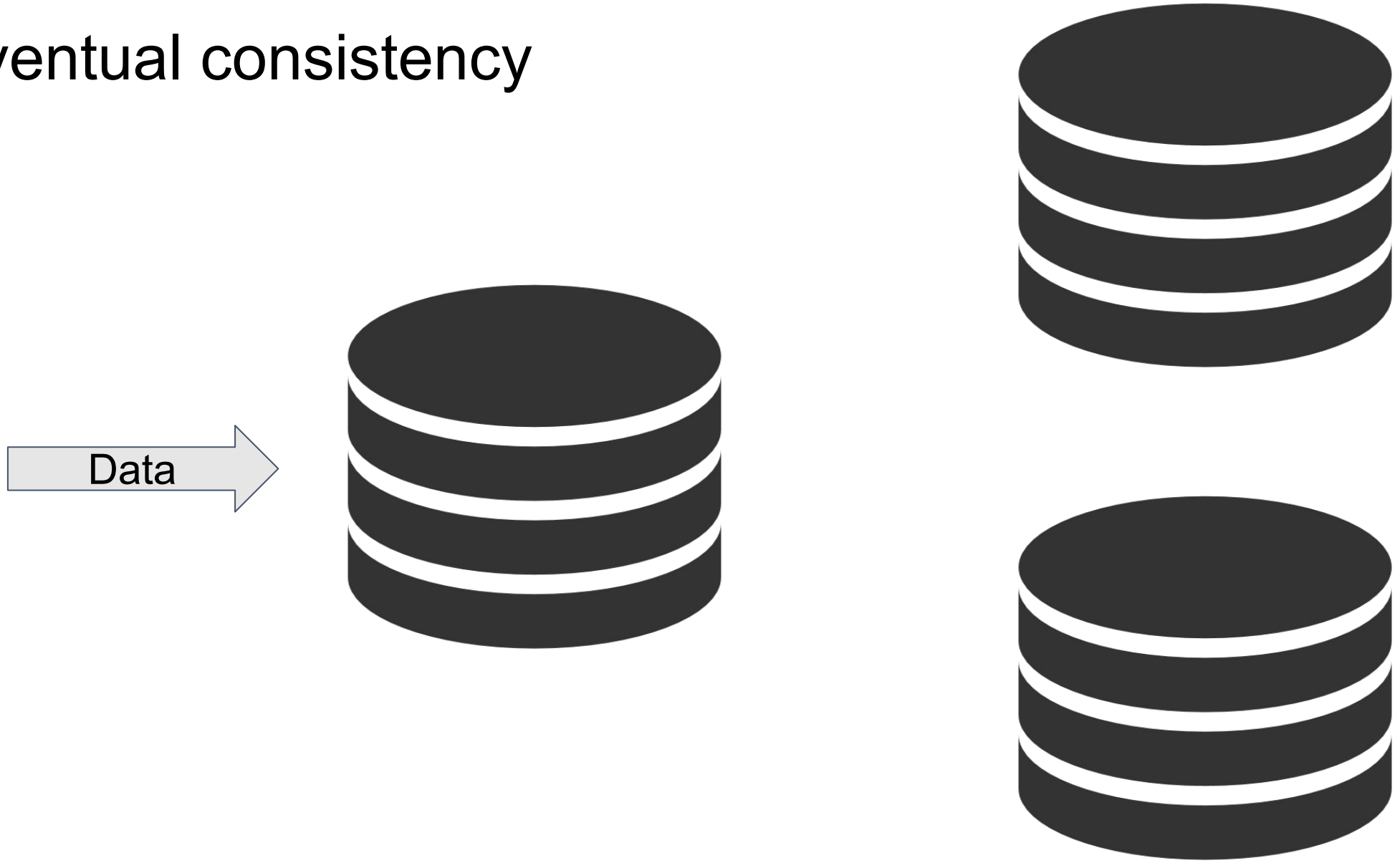
Авария



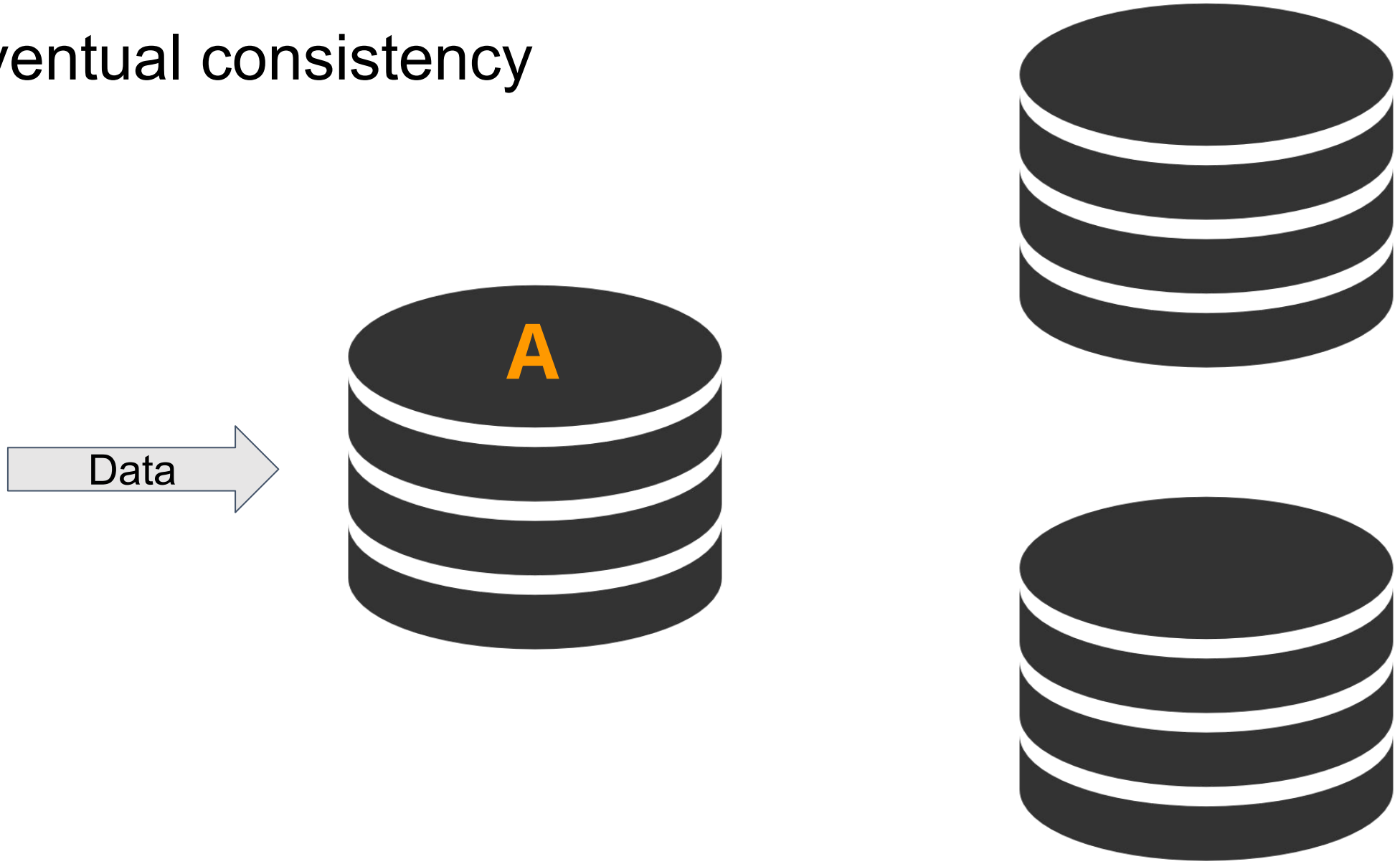
Просто добавь данных (опять)!



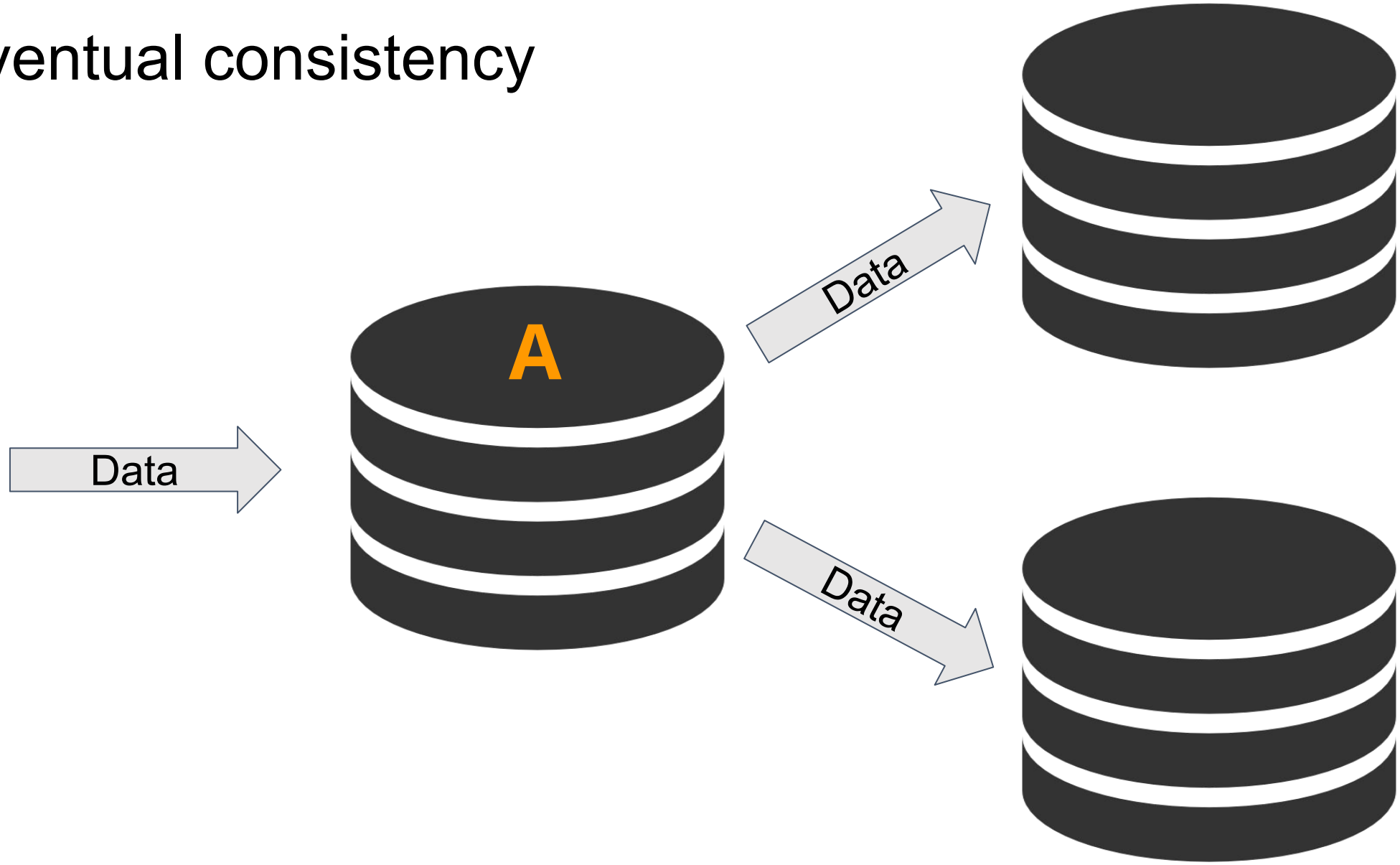
Eventual consistency



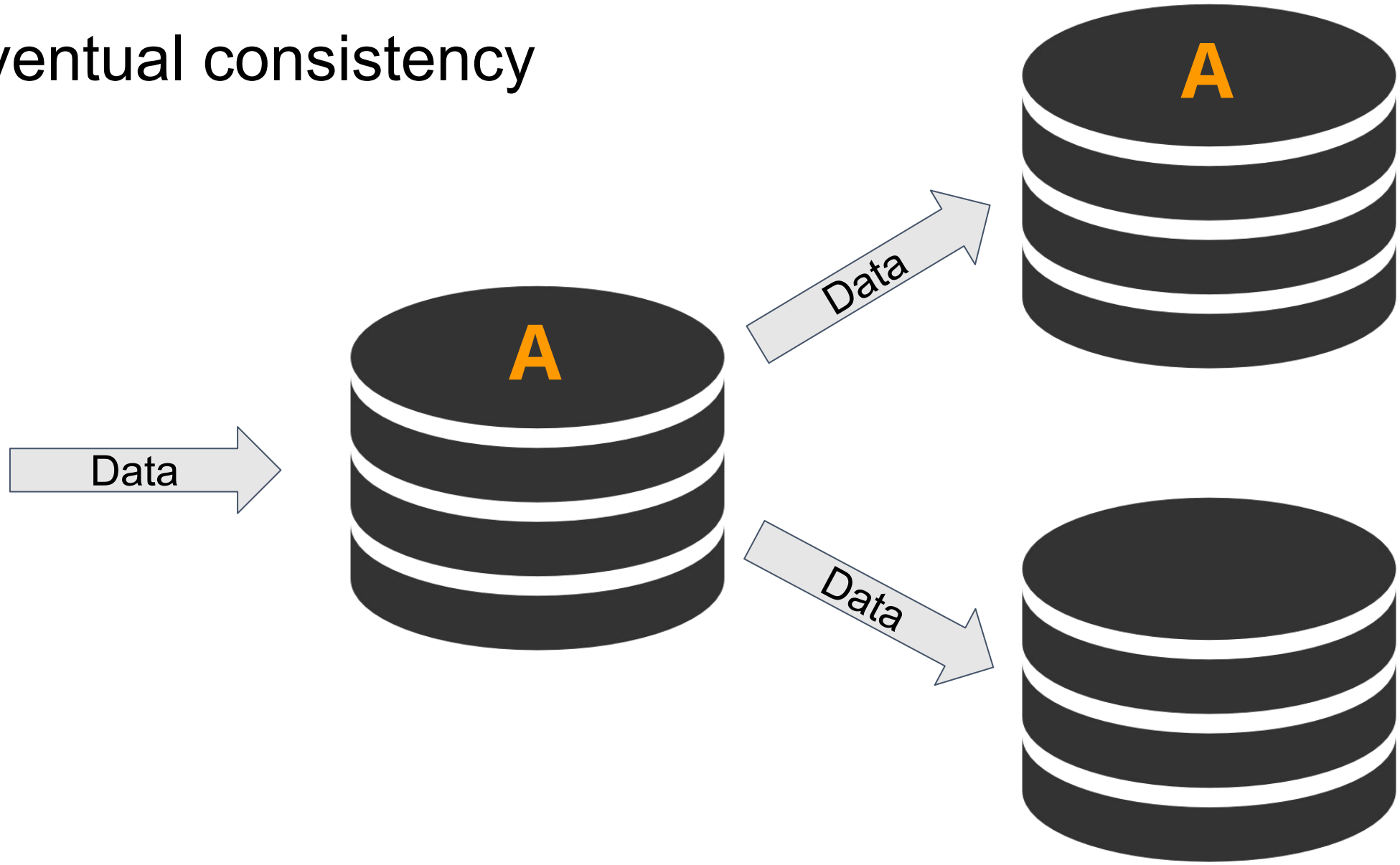
Eventual consistency



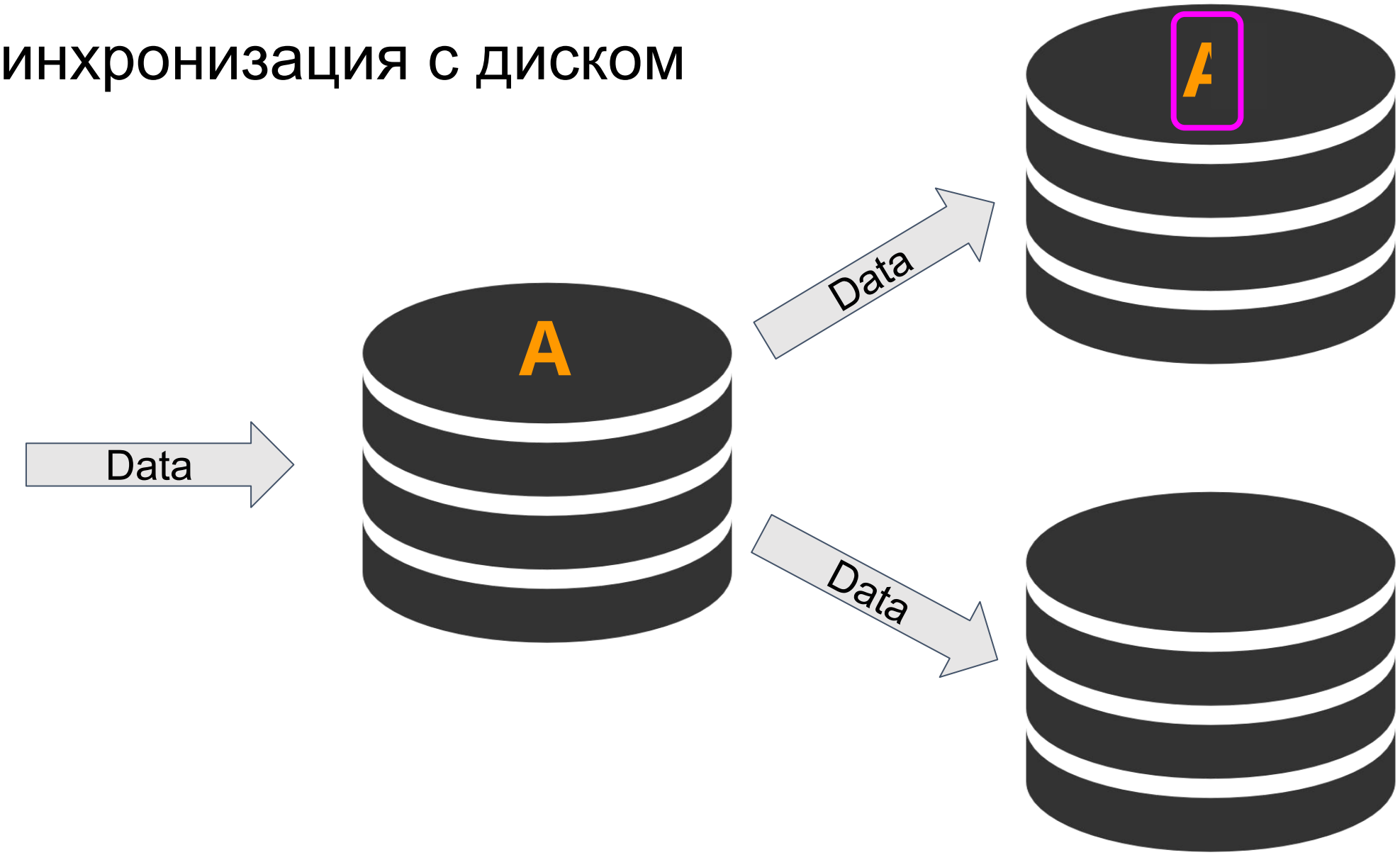
Eventual consistency



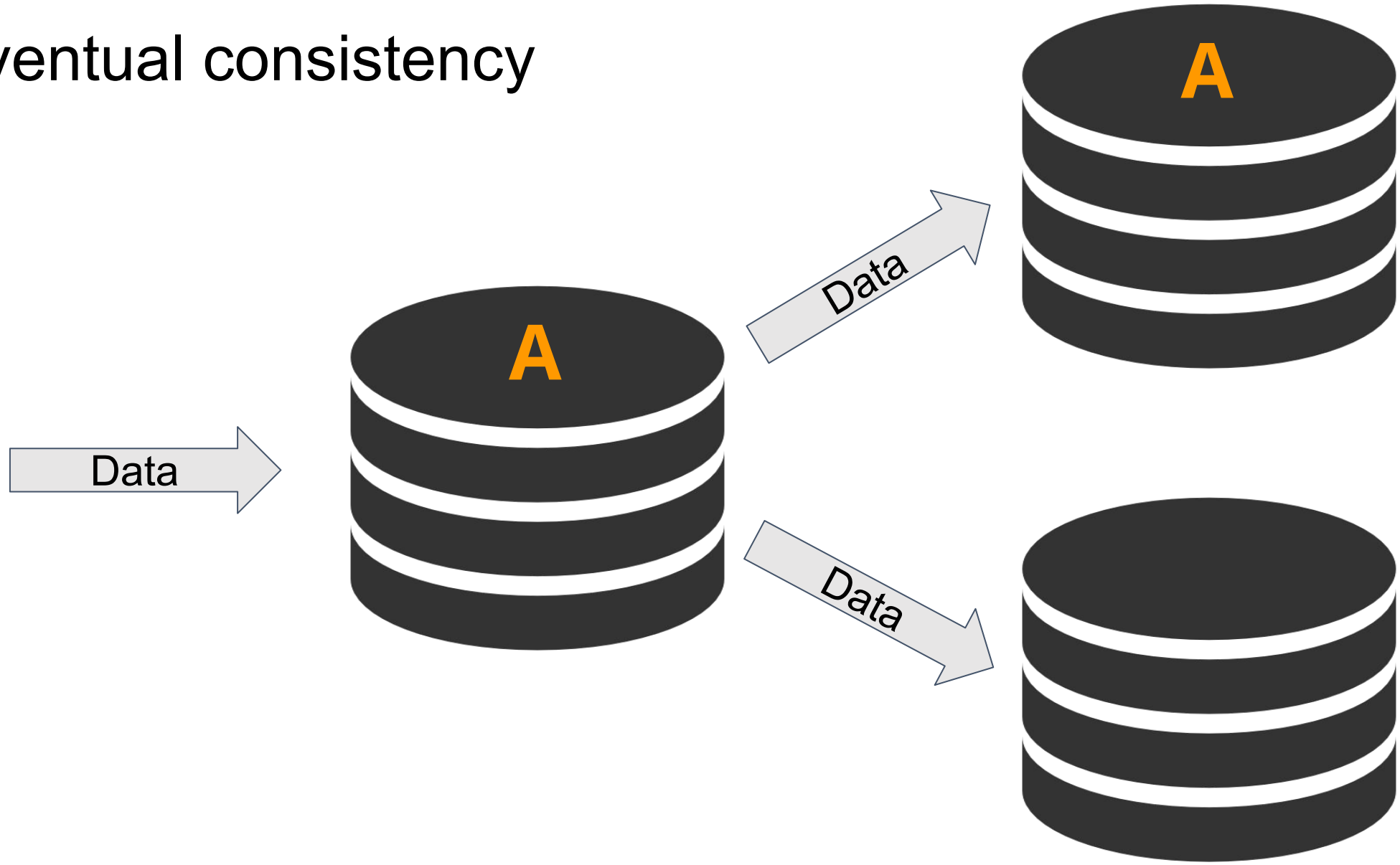
Eventual consistency



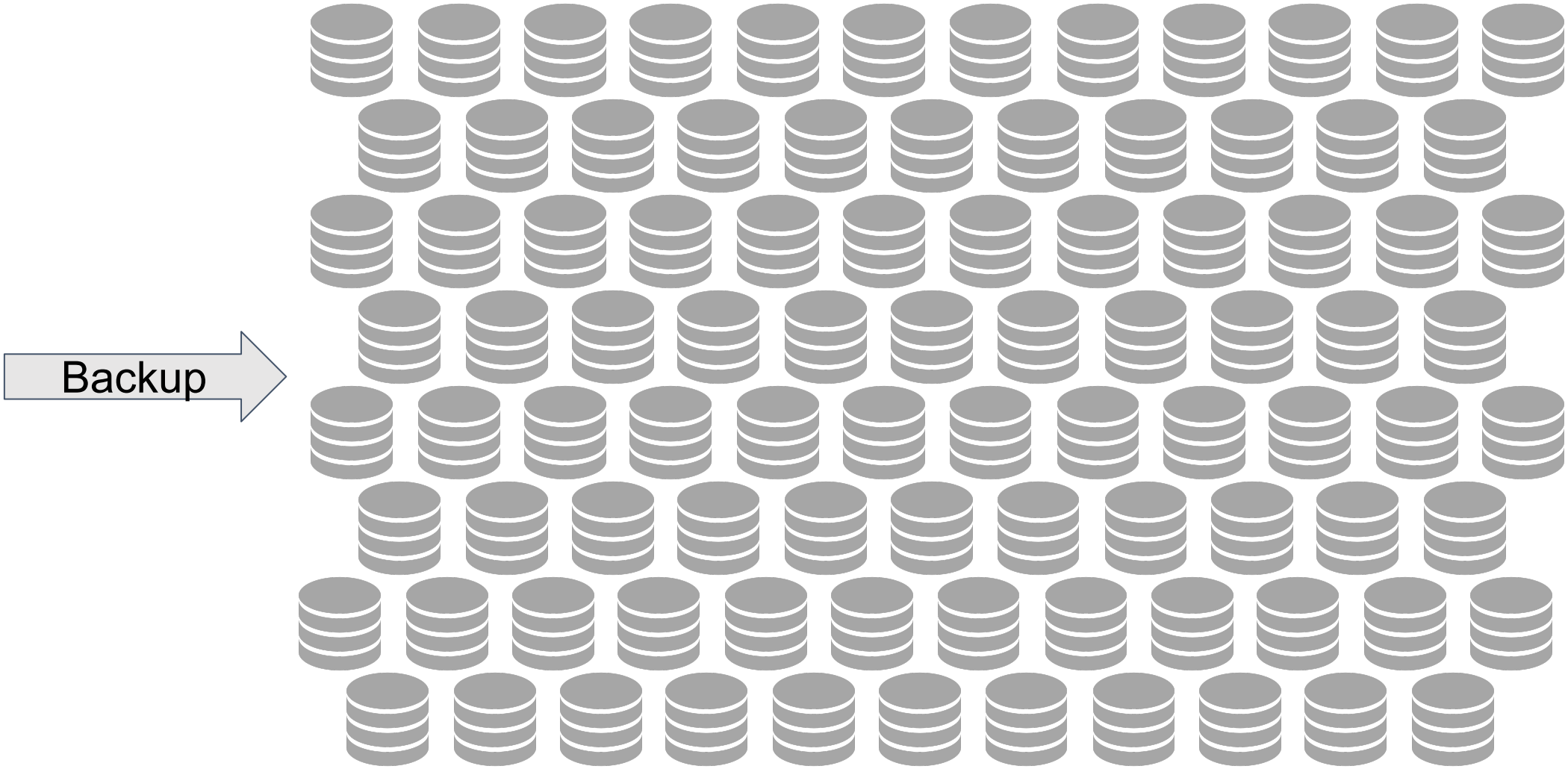
Синхронизация с диском



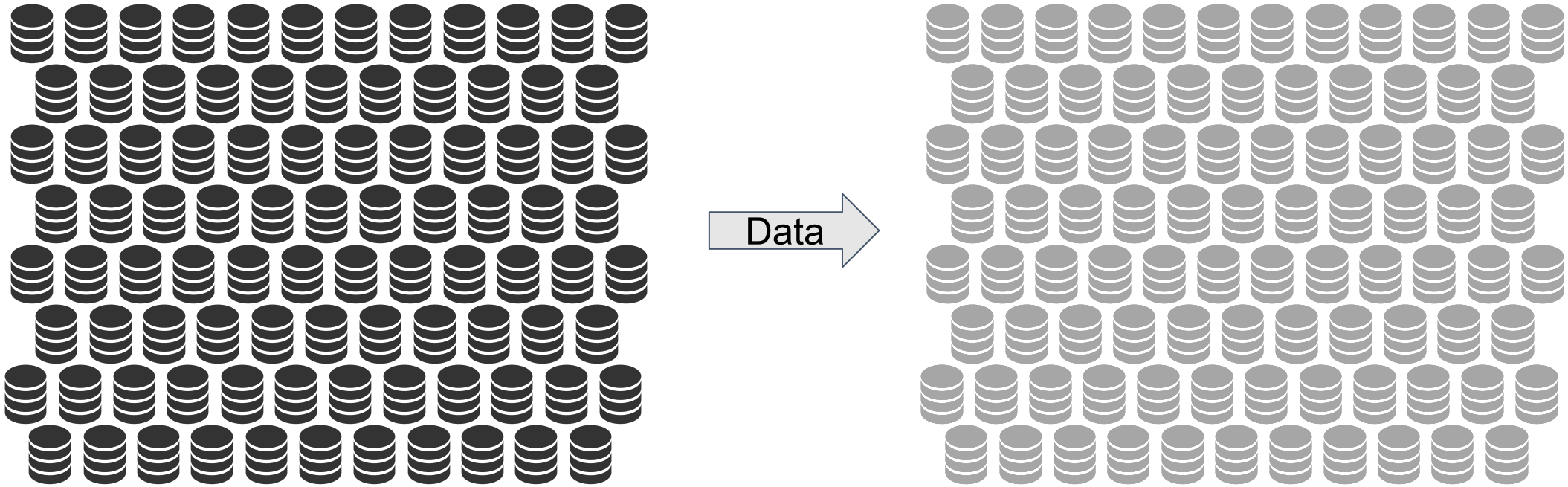
Eventual consistency



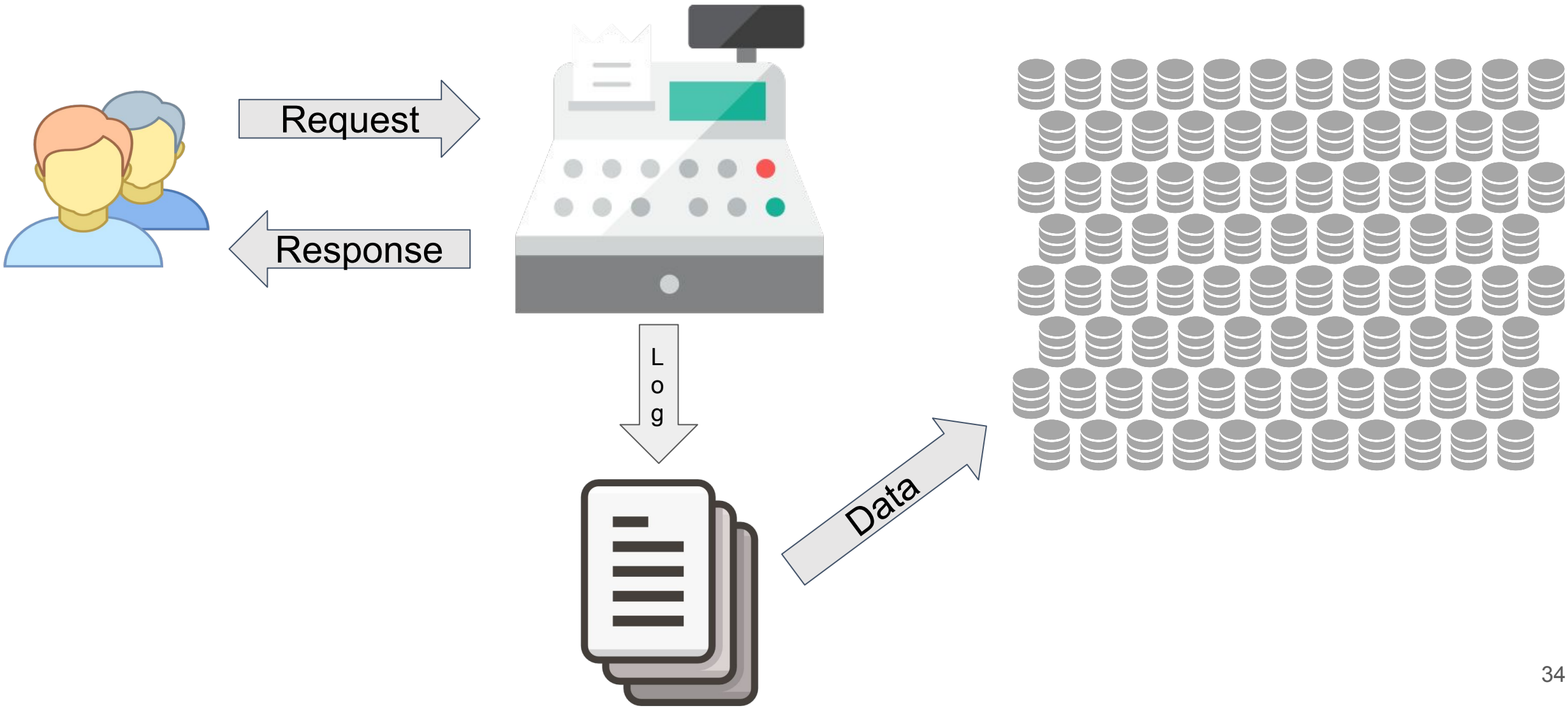
#1 Восстанавливаем из бэкапа



#2 Восстанавливаем из резервного кластера

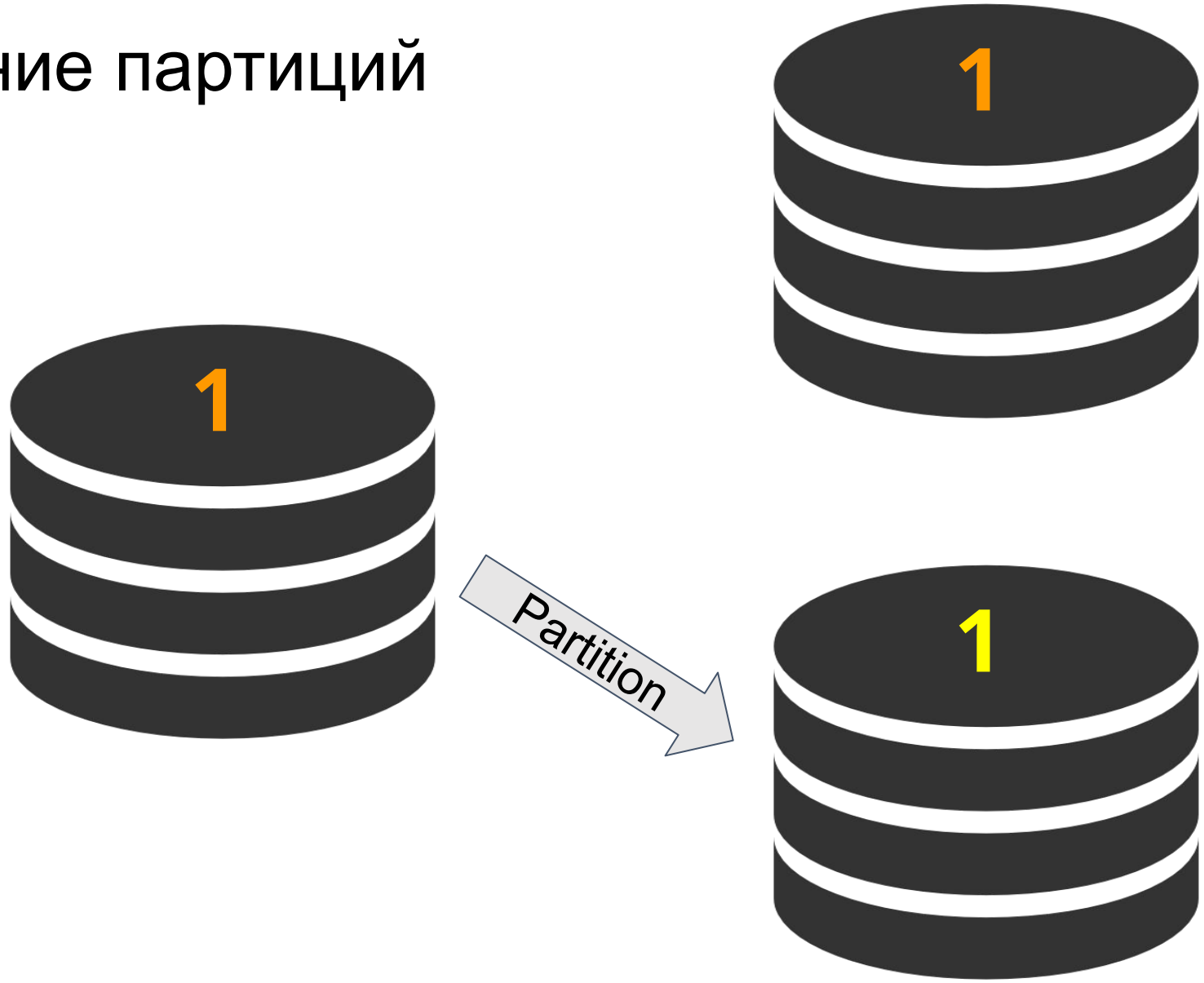


#3 Восстанавливаем из прикладных журналов



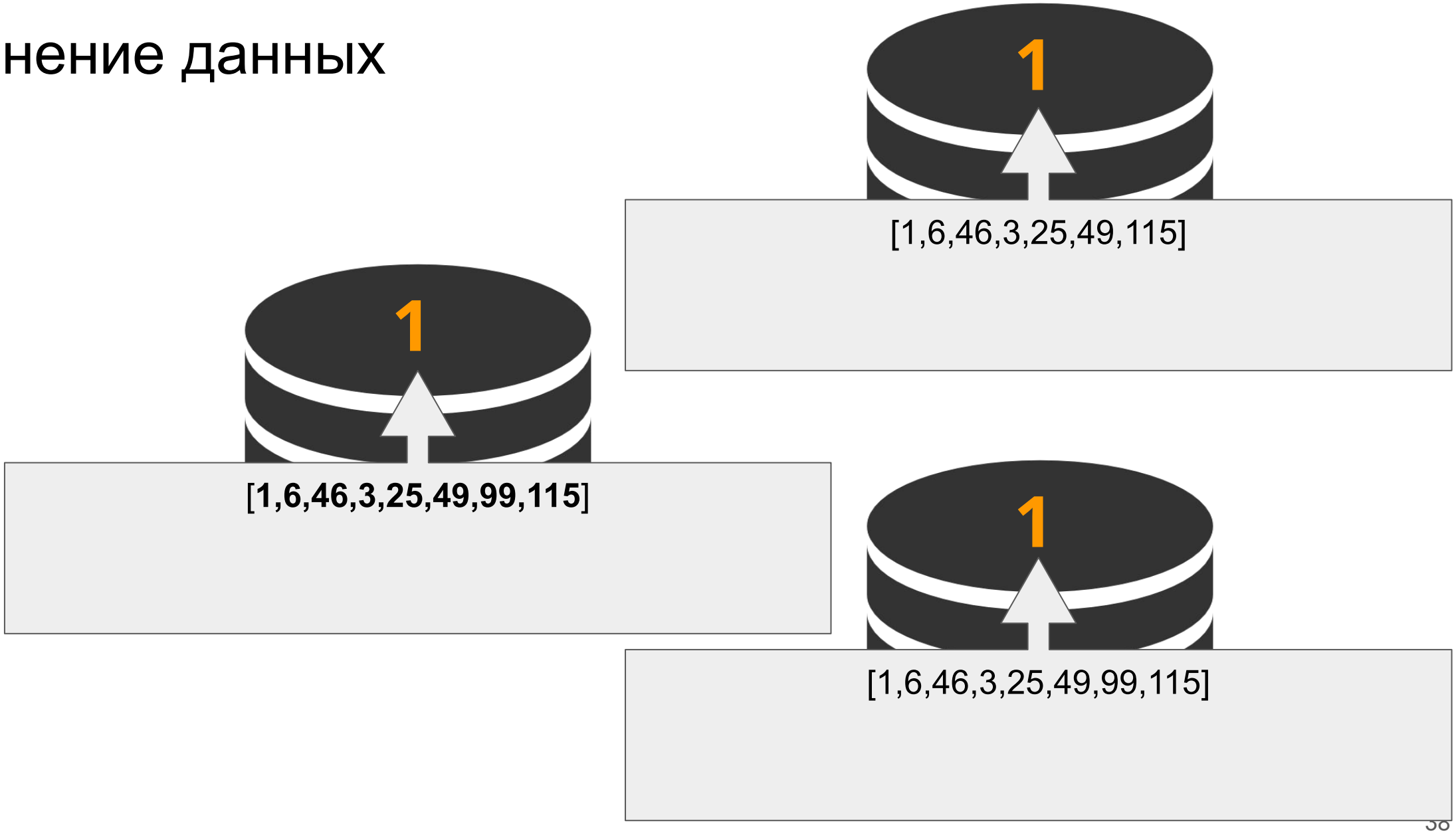
Eventual consistency

Восстановление партиций

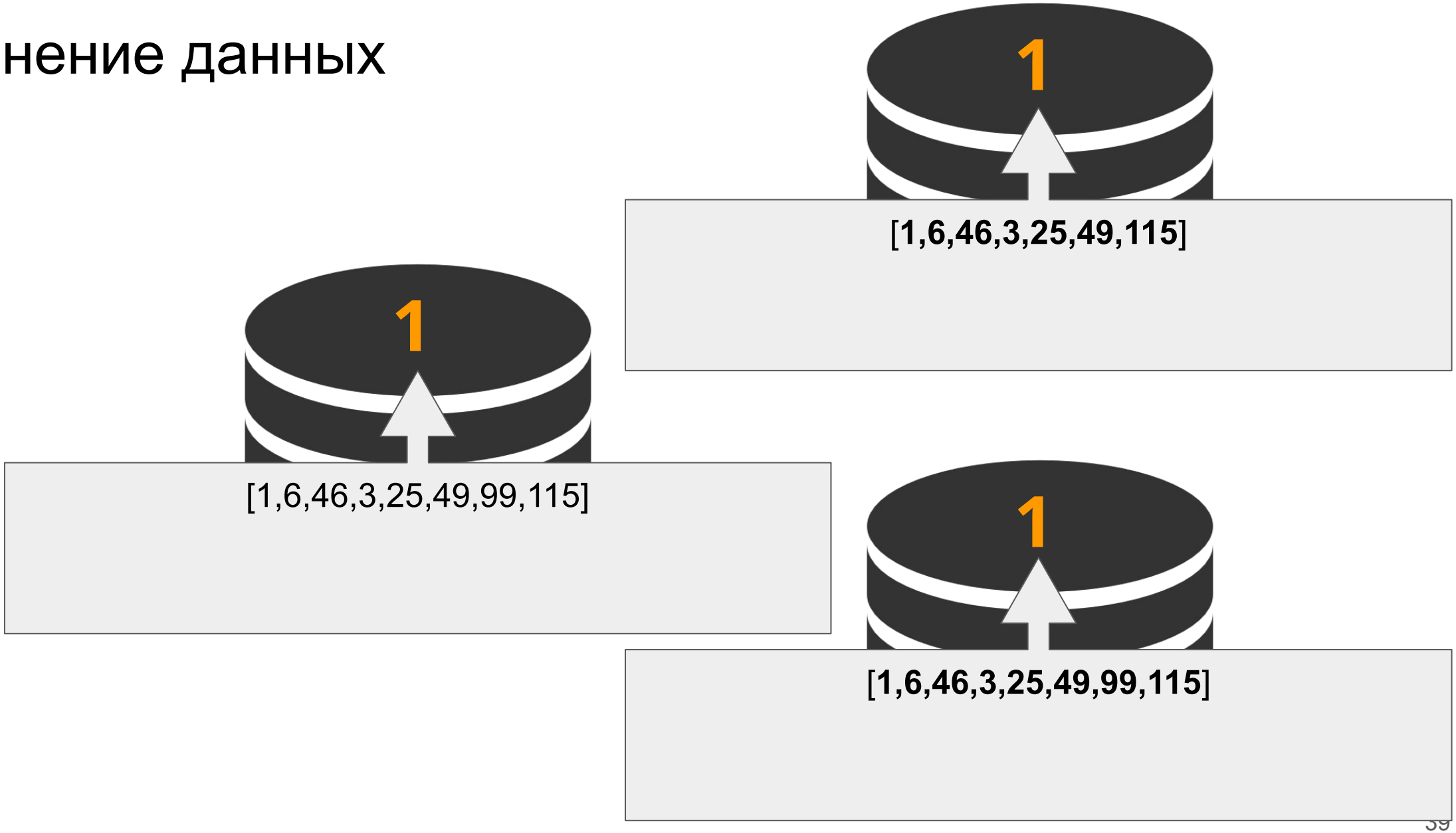


Восстановление партиций

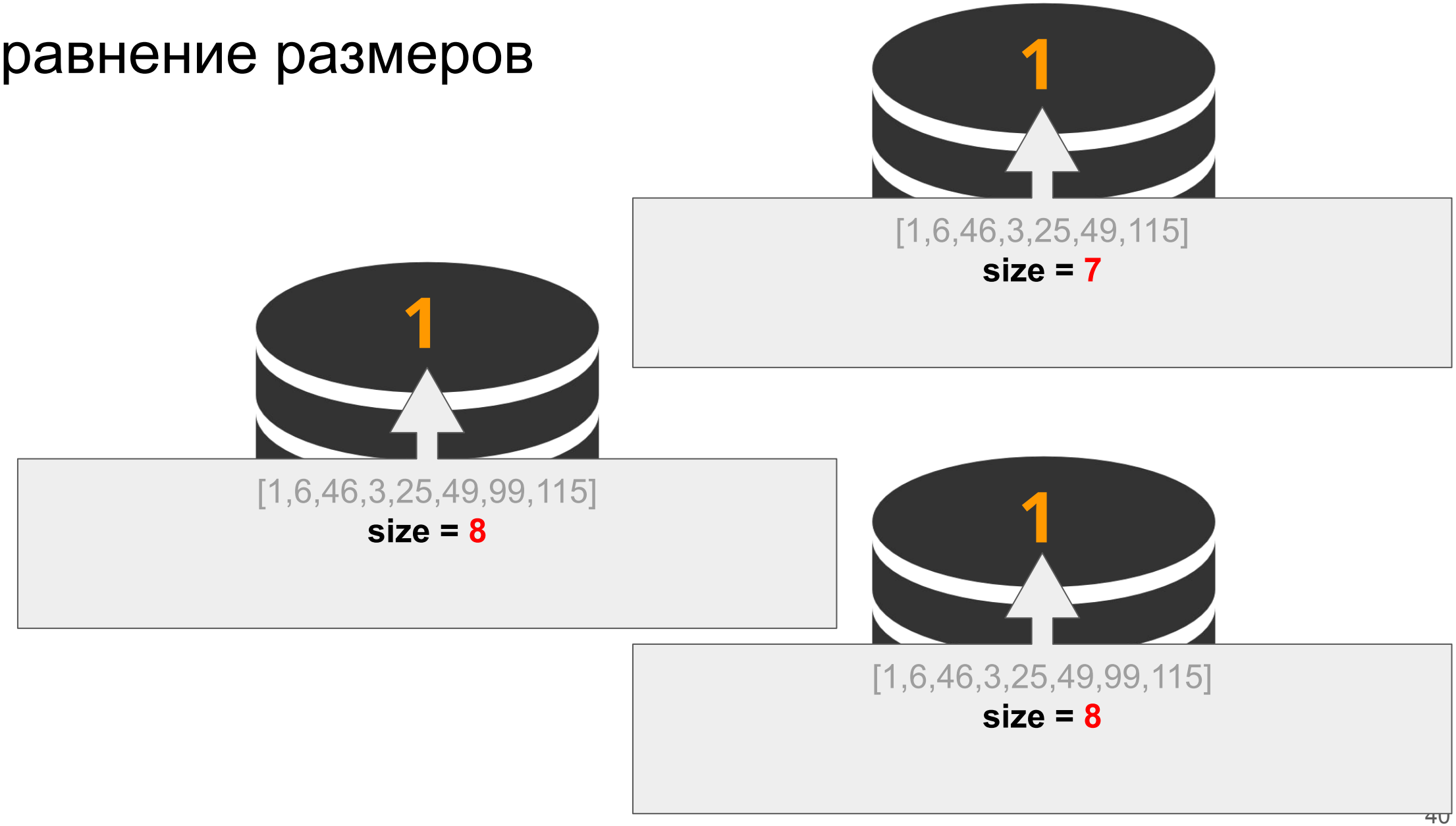
Сравнение данных



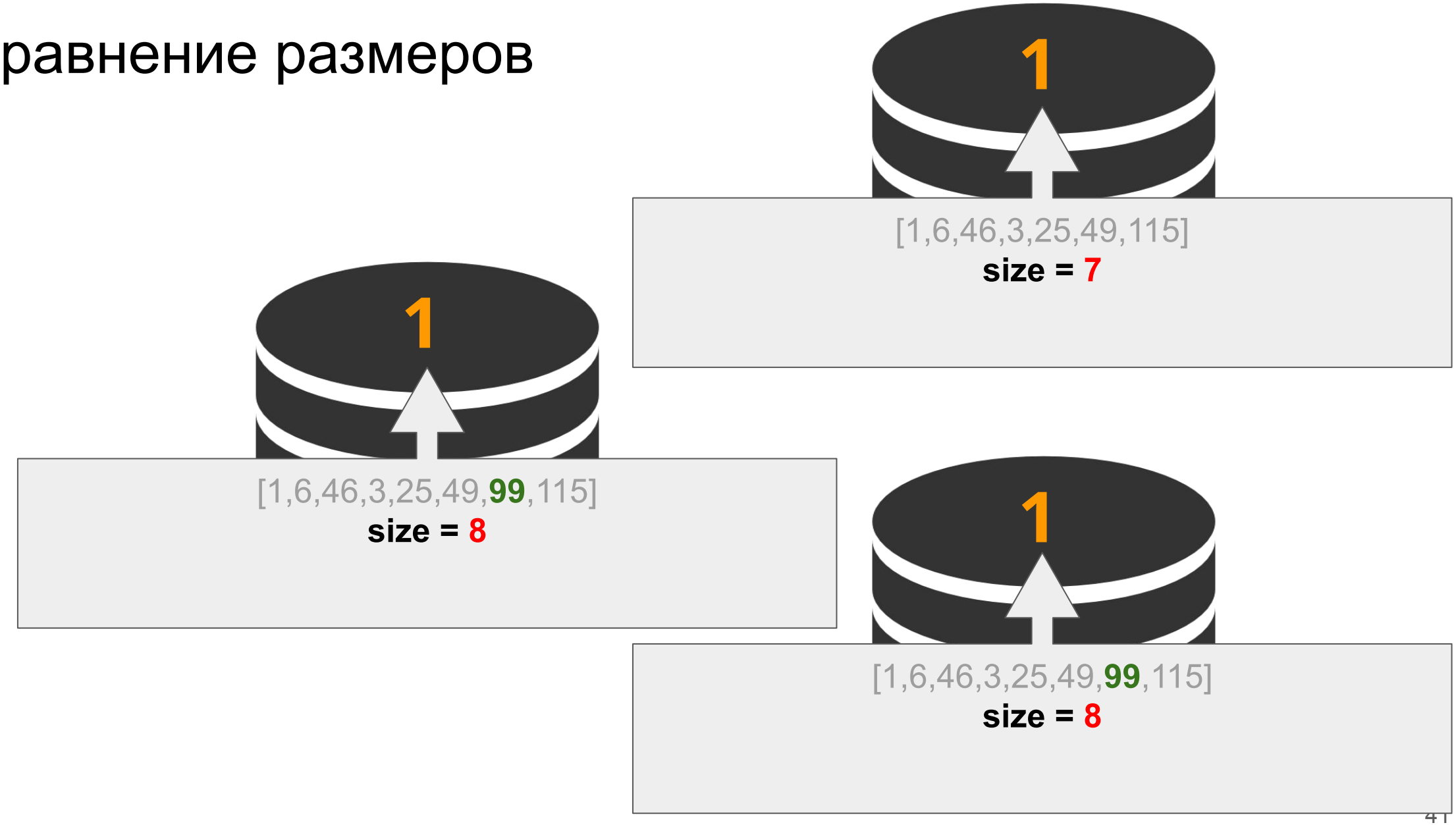
Сравнение данных



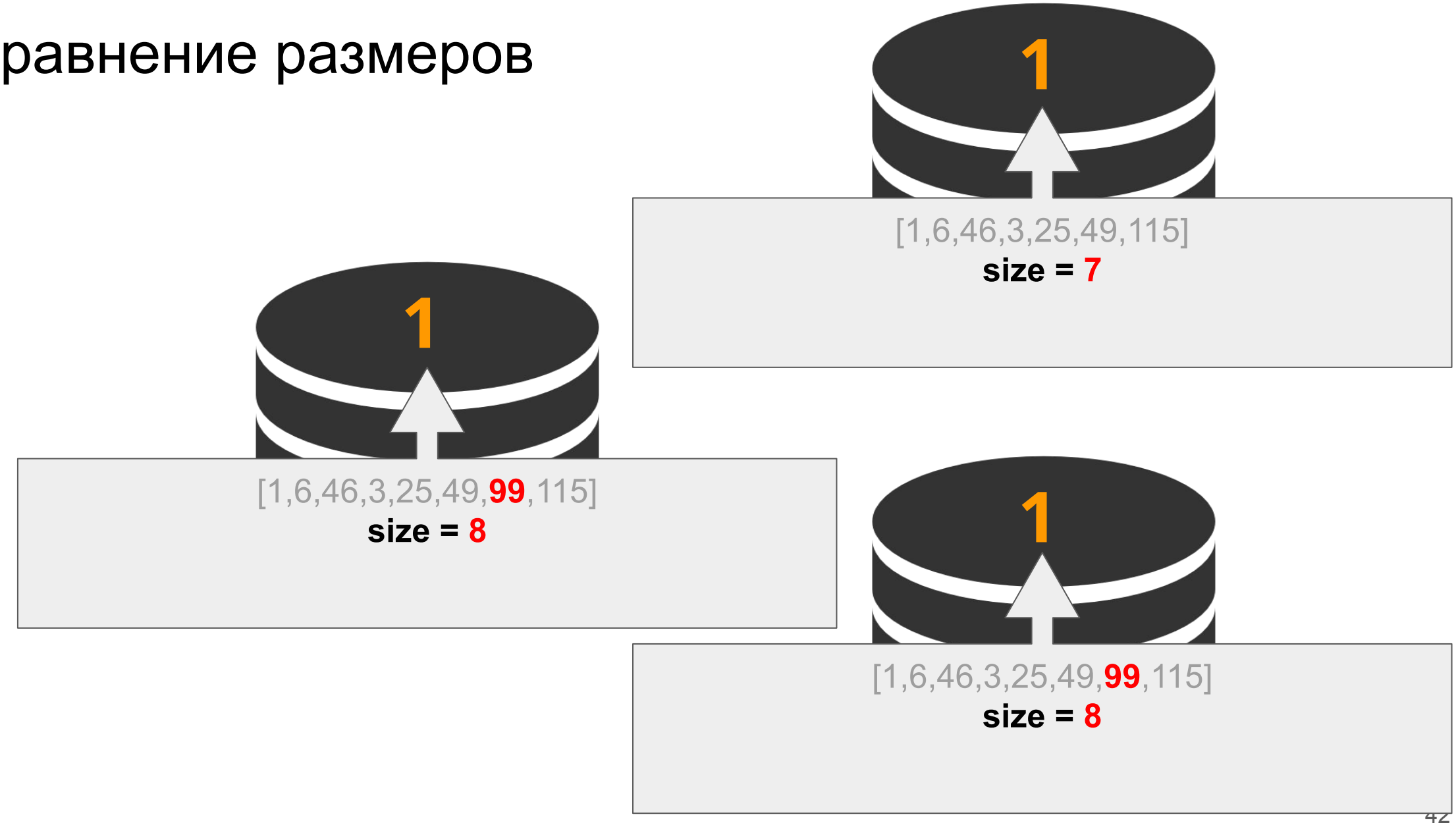
#1 Сравнение размеров



#1 Сравнение размеров

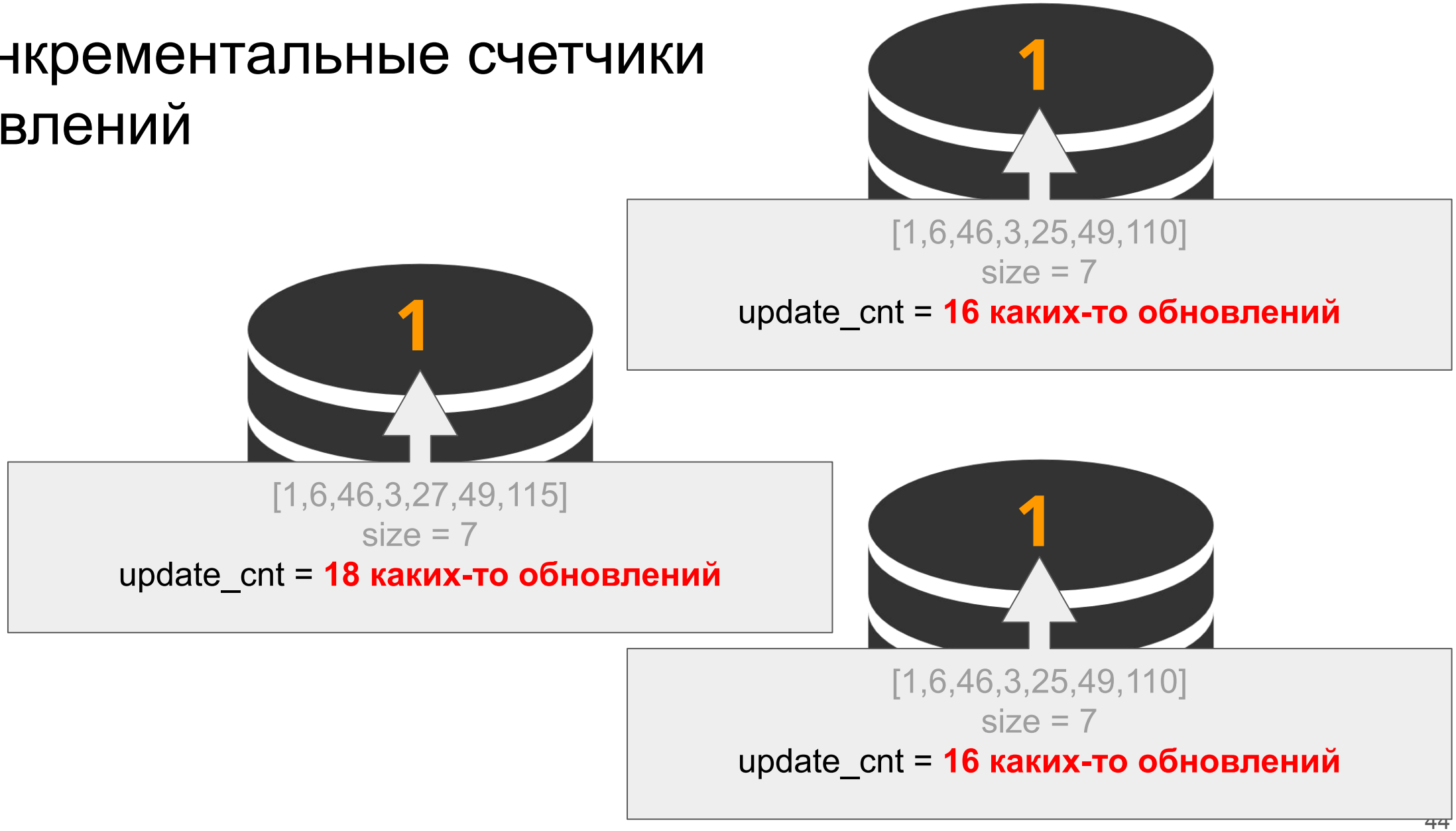


#1 Сравнение размеров

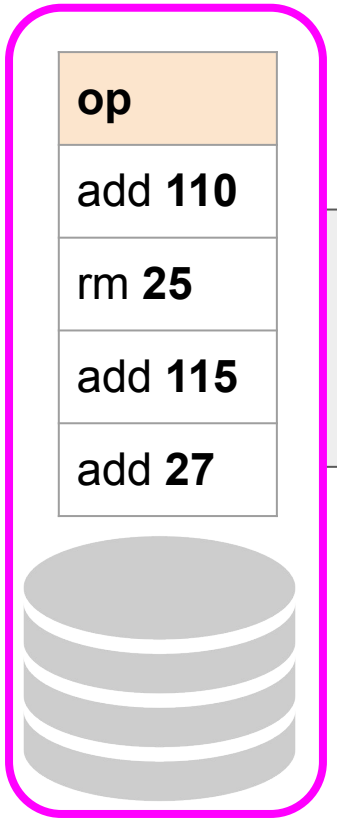


#2 Инкрементальные счетчики обновлений

#2 Инкрементальные счетчики обновлений



#2 Инкрементальные счетчики обновлений



было [1,6,46,3,25,49] -> **стало** [1,6,46,3,25,49,**110**]
size = 7
update_cnt = 15 -> 16

было [1,6,46,3,**25**,49] -> **стало** [1,6,46,3,**27**,49,**115**]
size = 7
update_cnt = 15 -> 18

было [1,6,46,3,25,49] -> **стало** [1,6,46,3,25,49,**110**]
size = 7
update_cnt = 15 -> 16

#2 Инкрементальные счетчики обновлений

op
add **110**
rm **25**
add **115**
add **27**

cnt	op
16	rm **25**
17	add **115**
18	add **27**

было [1,6,46,3,**25**,49] -> **стало** [1,6,46,3,**27**,49,**115**]
size = 7
update_cnt = **15** -> **18**

было [1,6,46,3,25,49] -> **стало** [1,6,46,3,25,49,**110**]
size = 7
update_cnt = 15 -> 16

было [1,6,46,3,25,49] -> **стало** [1,6,46,3,25,49,**110**]
size = 7
update_cnt = 15 -> 16

#2 Инкрементальные счетчики обновлений

op
add **110**
rm **25**
add **115**
add **27**

cnt	op
16	rm **25**
17	add **115**
18	add **27**

было [1,6,46,3,**25**,49] -> **стало** [1,6,46,3,**27**,49,**115**]
size = 7
update_cnt = **15** -> **18**

было [1,6,46,3,25,49] -> **стало** [1,6,46,3,25,49,**110**]
size = 7
update_cnt = **15** -> **16**

cnt	op
16	add **110**

было [1,6,46,3,25,49] -> **стало** [1,6,46,3,25,49,**110**]
size = 7
update_cnt = **15** -> **16**

#2 Инкрементальные счетчики обновлений

op
add **110**
rm **25**
add **115**
add **27**

cnt	op
16	rm **25**
17	add **115**
18	add **27**

было [1,6,46,3,**25**,49] -> **стало** [1,6,46,3,**27**,49,**115**]
size = 7
update_cnt = **15** -> **18**

было [1,6,46,3,25,49] -> **стало** [1,6,46,3,25,49,**110**]
size = 7
update_cnt = **15** -> **16**

cnt	op
16	add **110**

было [1,6,46,3,25,49] -> **стало** [1,6,46,3,25,49,**110**]
size = 7
update_cnt = **15** -> **16**

#2 Инкрементальные счетчики обновлений

op
add **110**
rm **25**
add **115**
add **27**

cnt	op
16	rm **25**
17	add **115**
18	add **27**

было [1,6,46,3,**25**,49] -> **стало** [1,6,46,3,**27**,49,**115**]
size = 7
update_cnt = **18**

было [1,6,46,3,**25**,49] -> **стало** [1,6,46,3,**27**,49,**110**]
size = 7
update_cnt = **18**

cnt	op
16	add **110**
17	rm **25**
18	add **27**

было [1,6,46,3,**25**,49] -> **стало** [1,6,46,3,**27**,49,**110**]
size = 7
update_cnt = **18**

#3 История обновлений

cnt	op
16	add **110**
17	rm **25**
18	add **115**
19	add **27**

cnt	op
17	rm **25**
18	add **115**
19	add **27**

[1,6,46,3,27,49,115]
size = 7
update_cnt = **1, 2, 3, 4 … 15, 17, 18, 19**

[1,6,46,3,25,49,110]
size = 7
update_cnt = **1, 2, 3, 4 … 15, 16**

cnt	op
16	add **110**

[1,6,46,3,25,49,110]
size = 7
update_cnt = **1, 2, 3, 4 … 15, 16**

#3 История обновлений

cnt	op
16	add **110**
17	rm **25**
18	add **115**
19	add **27**

cnt	op
17	rm **25**
18	add **115**
19	add **27**

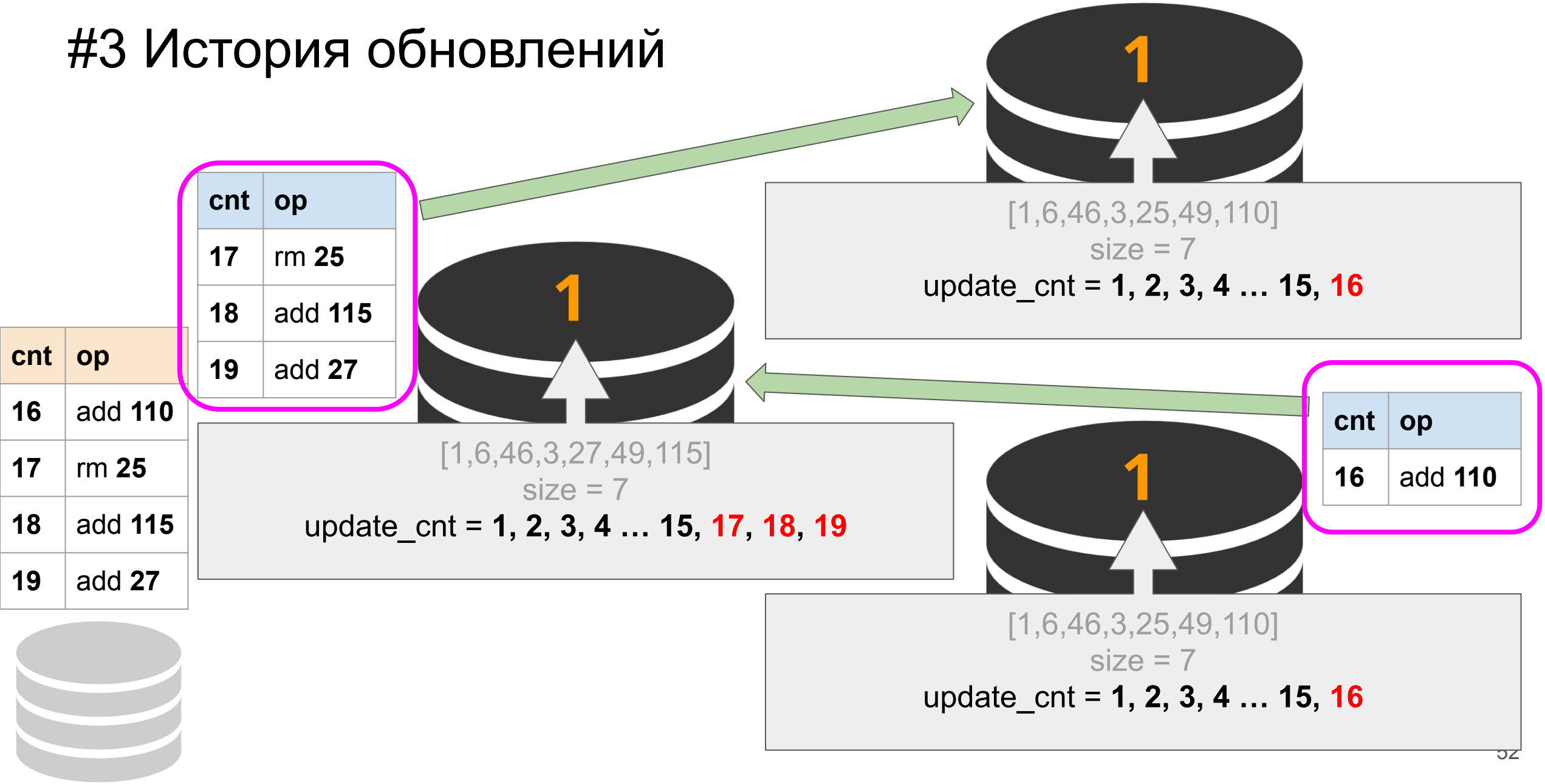
[1,6,46,3,27,49,115]
size = 7
update_cnt = **1, 2, 3, 4 … 15, 17, 18, 19**

[1,6,46,3,25,49,110]
size = 7
update_cnt = **1, 2, 3, 4 … 15, 16**

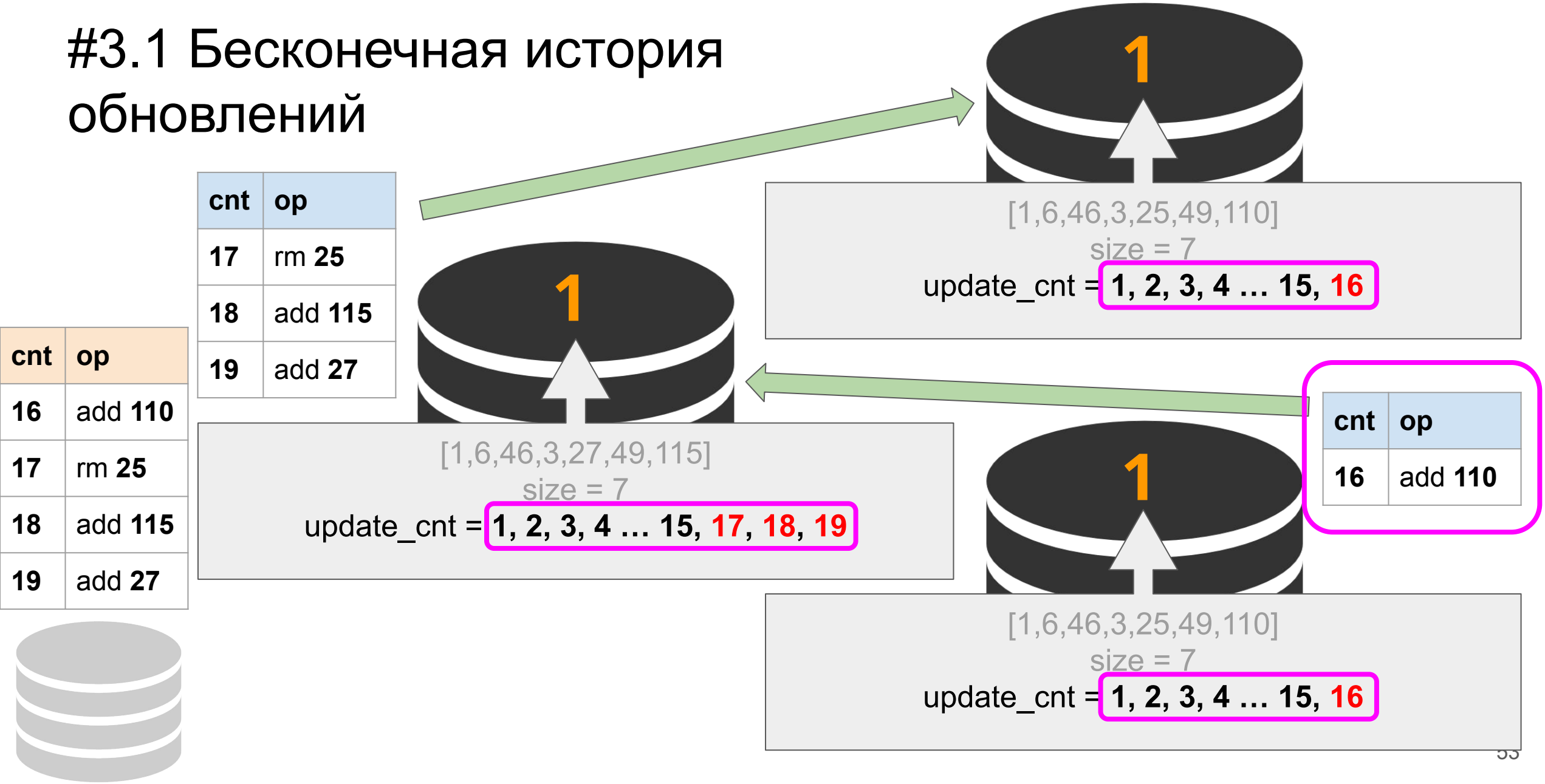
cnt	op
16	add **110**

[1,6,46,3,25,49,110]
size = 7
update_cnt = **1, 2, 3, 4 … 15, 16**

#3 История обновлений



#3.1 Бесконечная история обновлений



#3.2 Компактная история обновлений

cnt	op
16	add **110**
17	rm **25**
18	add **115**
19	add **27**

cnt	op
17	rm **25**
18	add **115**
19	add **27**

[1,6,46,3,27,49,115]
size = 7
update_cnt = [LWM=**15**, missed=[**16**], HWM=**19**]

[1,6,46,3,25,49,110]
size = 7
update_cnt = [LWM=**16** missed=[], HWM=**16**]

cnt	op
16	add **110**

[1,6,46,3,25,49,110]
size = 7
update_cnt = [LWM=**16** missed=[], HWM=**16**]

#3.2 Компактная история обновлений

cnt	op
16	add **110**
17	rm **25**
18	add **115**
19	add **27**

cnt	op
17	rm **25**
18	add **115**
19	add **27**

[1,6,46,3,27,49,115]
size = 7
update_cnt = [LWM=**15**, missed=[**16**], HWM=**19**]

[1,6,46,3,25,49,110]
size = 7
update_cnt = [LWM=**16**, missed=[], HWM=**16**]

cnt	op
16	add **110**

[1,6,46,3,25,49,110]
size = 7
update_cnt = [LWM=**16**, missed=[], HWM=**16**]

#3.2 Компактная история обновлений

cnt	op
16	add **110**
17	rm **25**
18	add **115**
19	add **27**

cnt	op
17	rm **25**
18	add **115**
19	add **27**

[1,6,46,3,27,49,115]
size = 7
update_cnt = [LWM=**15**, missed=[**16**], HWM=**19**]

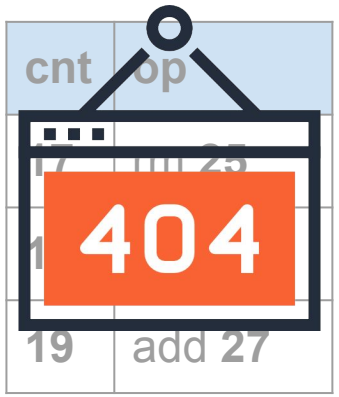
[1,6,46,3,25,49,110]
size = 7
update_cnt = [LWM=**16**, missed=[], HWM=**16**]

cnt	op
16	add **110**

[1,6,46,3,25,49,110]
size = 7
update_cnt = [LWM=**16**, missed=[], HWM=**16**]

#3.3 Утраченная история обновлений

cnt	op
16	add **110**
17	rm **25**
18	add **115**
19	add **27**



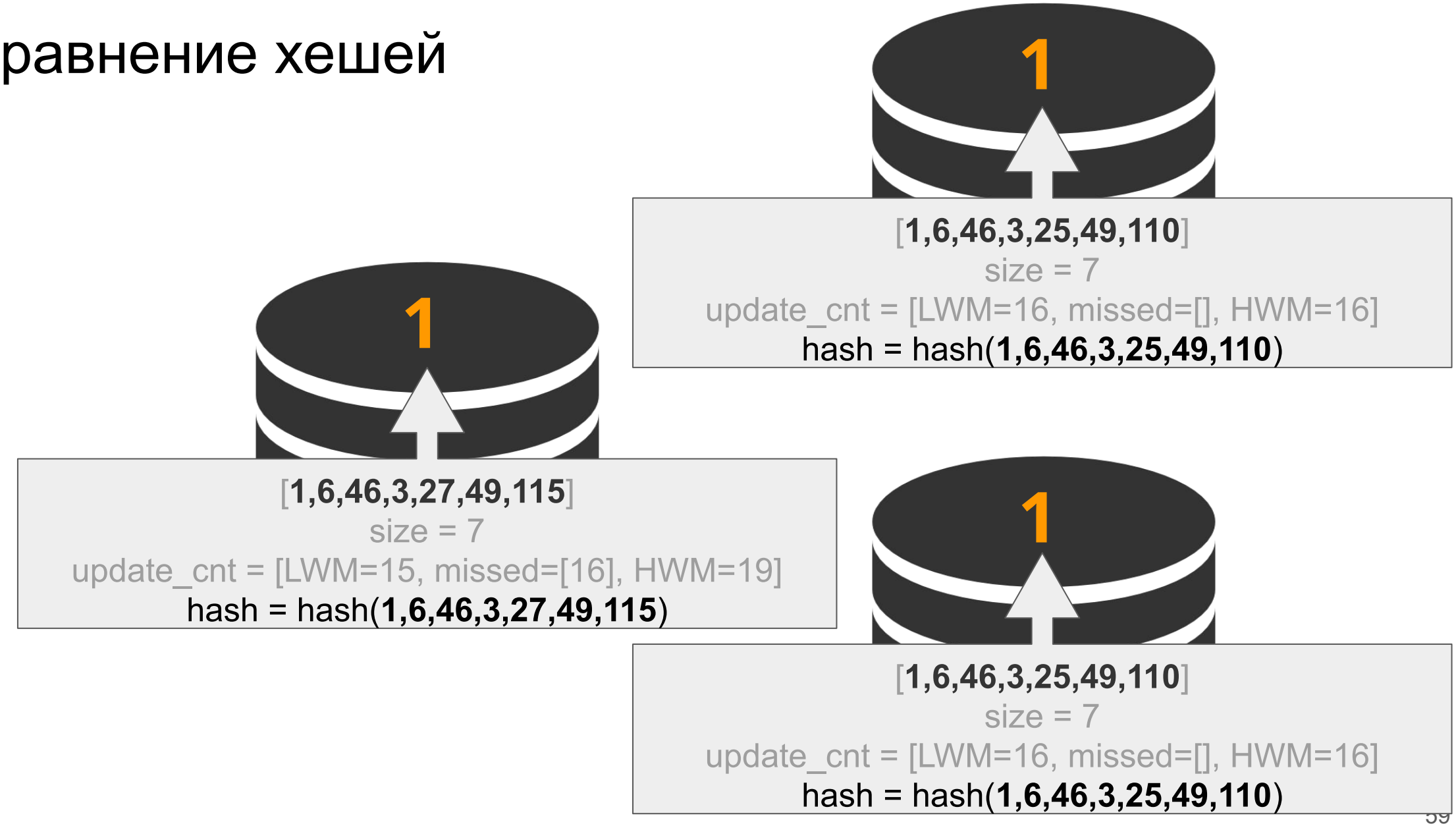
[1,6,46,3,27,49,115]
size = 7
update_cnt = [LWM=**15**, missed=[**16**], HWM=**19**]

[1,6,46,3,25,49,110]
size = 7
update_cnt = [LWM=**16**, missed=[], HWM=**16**]

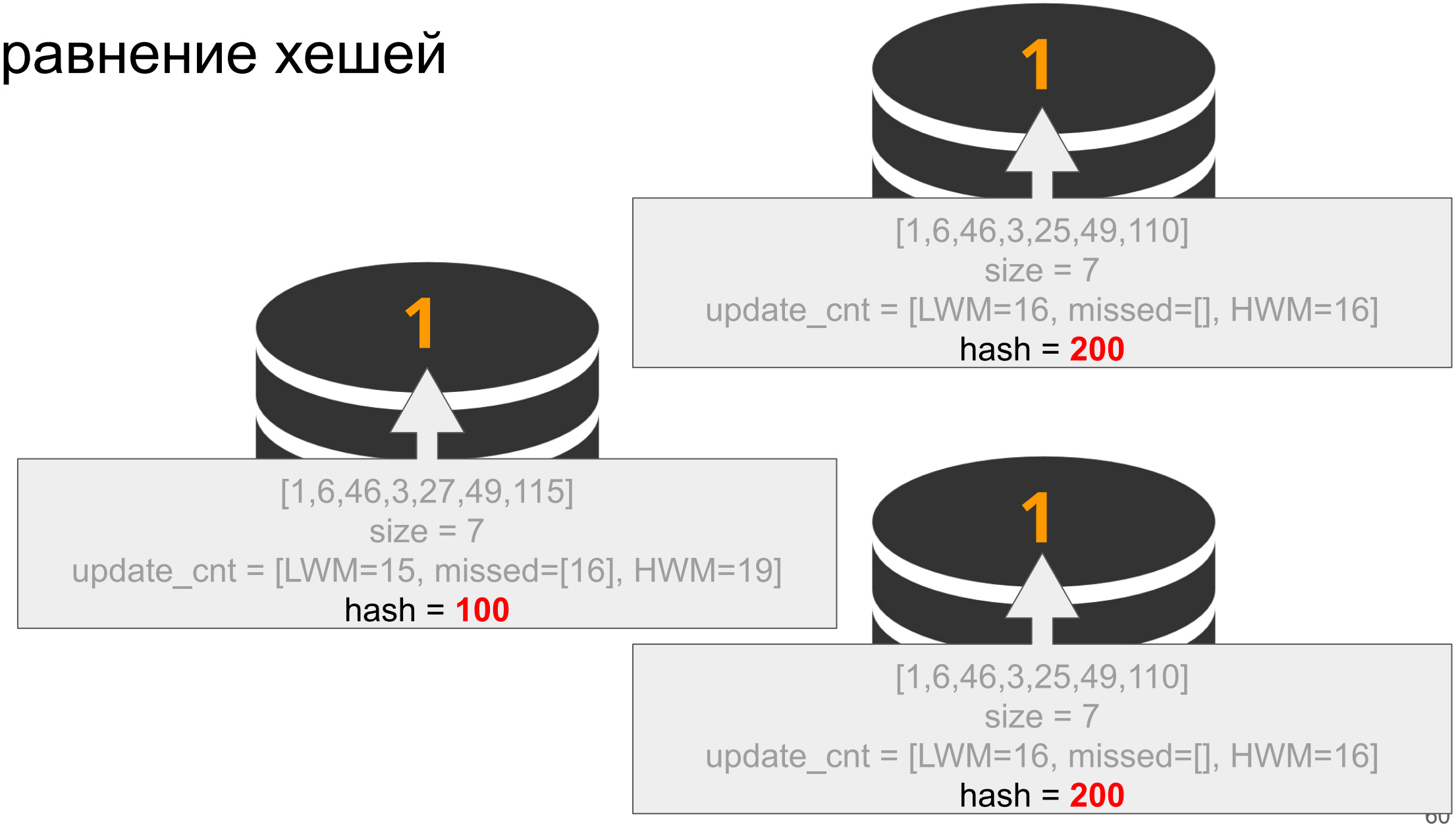
[1,6,46,3,25,49,110]
size = 7
update_cnt = [LWM=**16**, missed=[], HWM=**16**]

Вынужденное сравнение данных

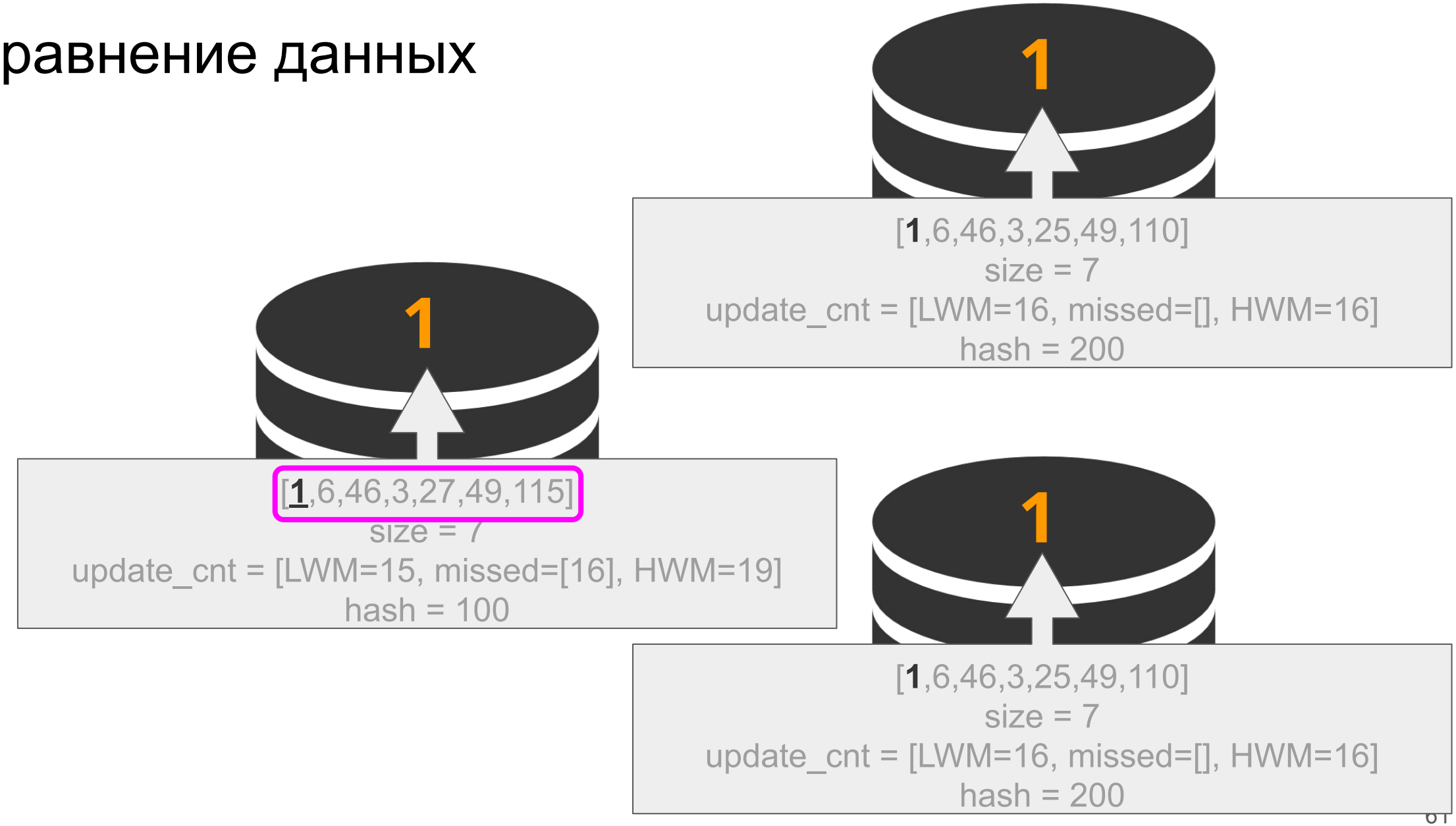
#4 Сравнение хешей



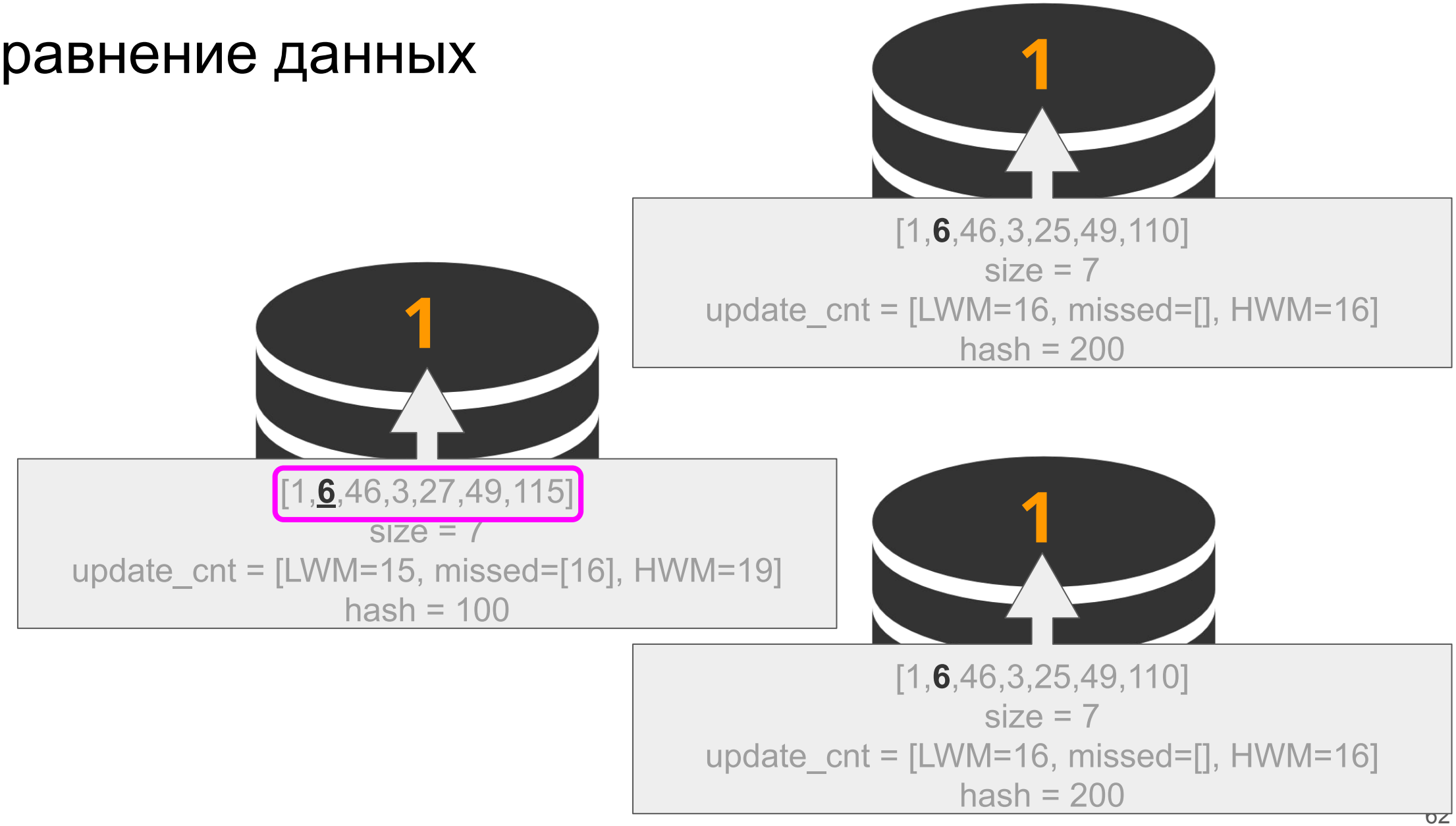
#4 Сравнение хешей



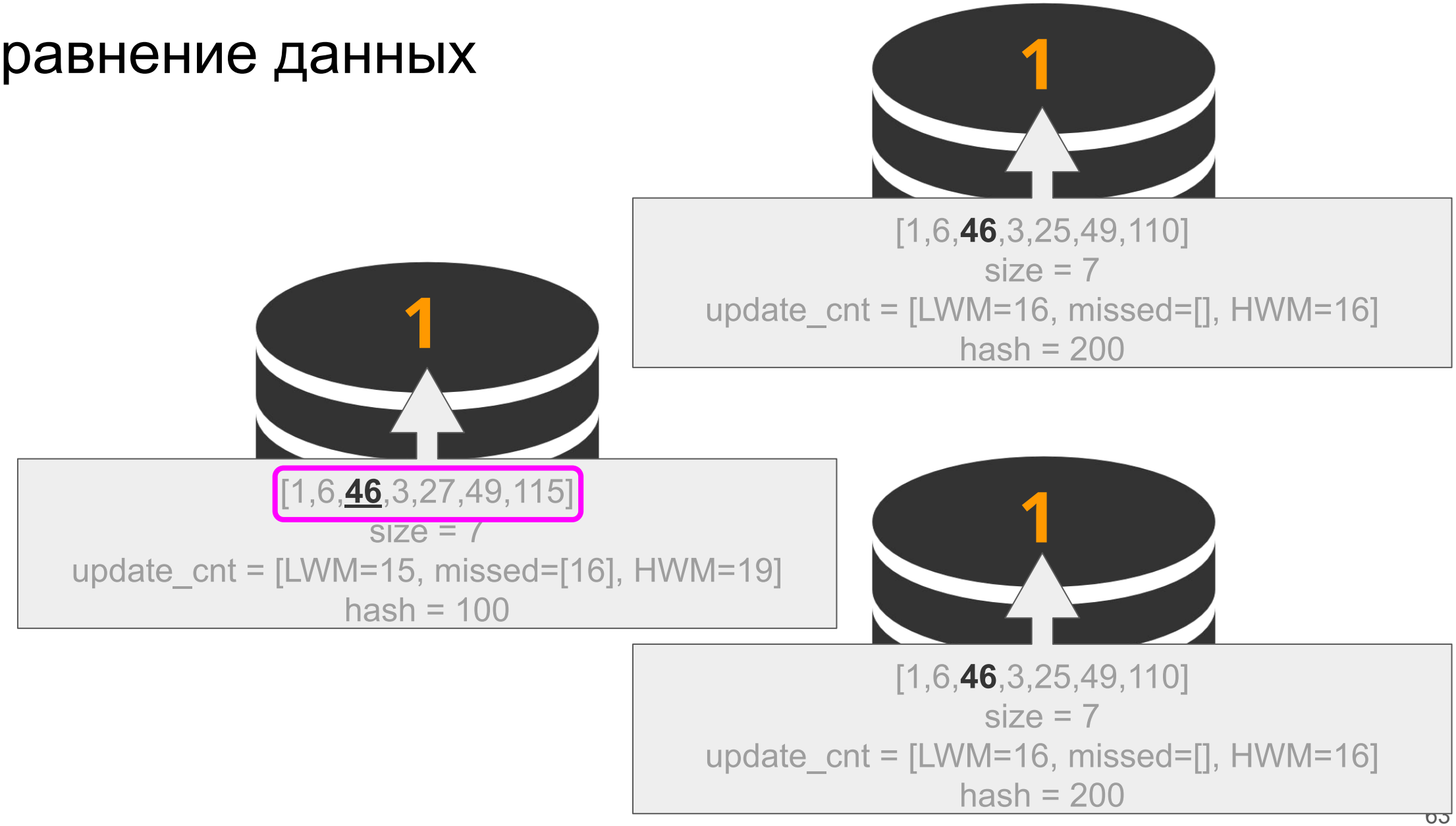
#5 Сравнение данных



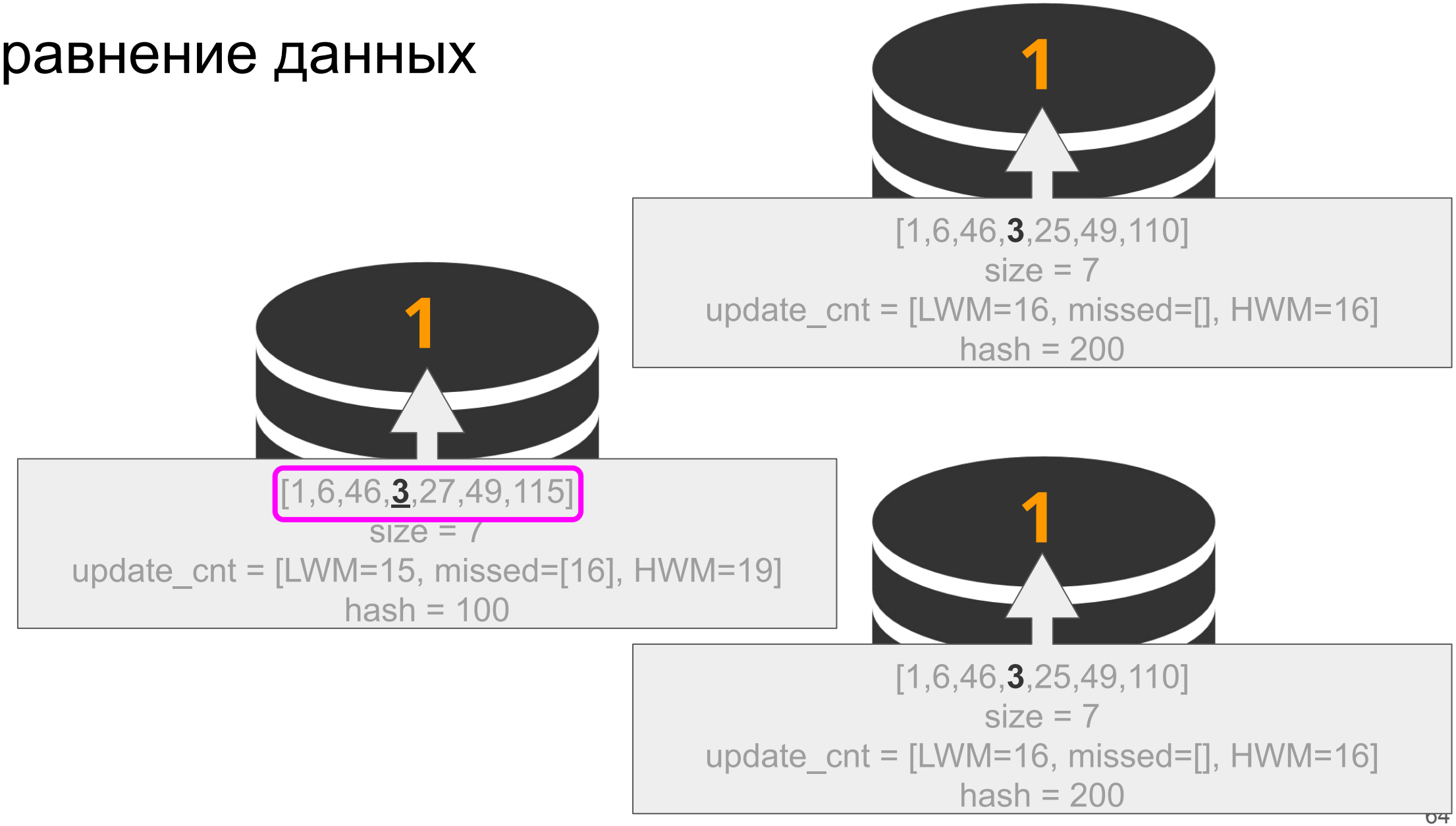
#5 Сравнение данных



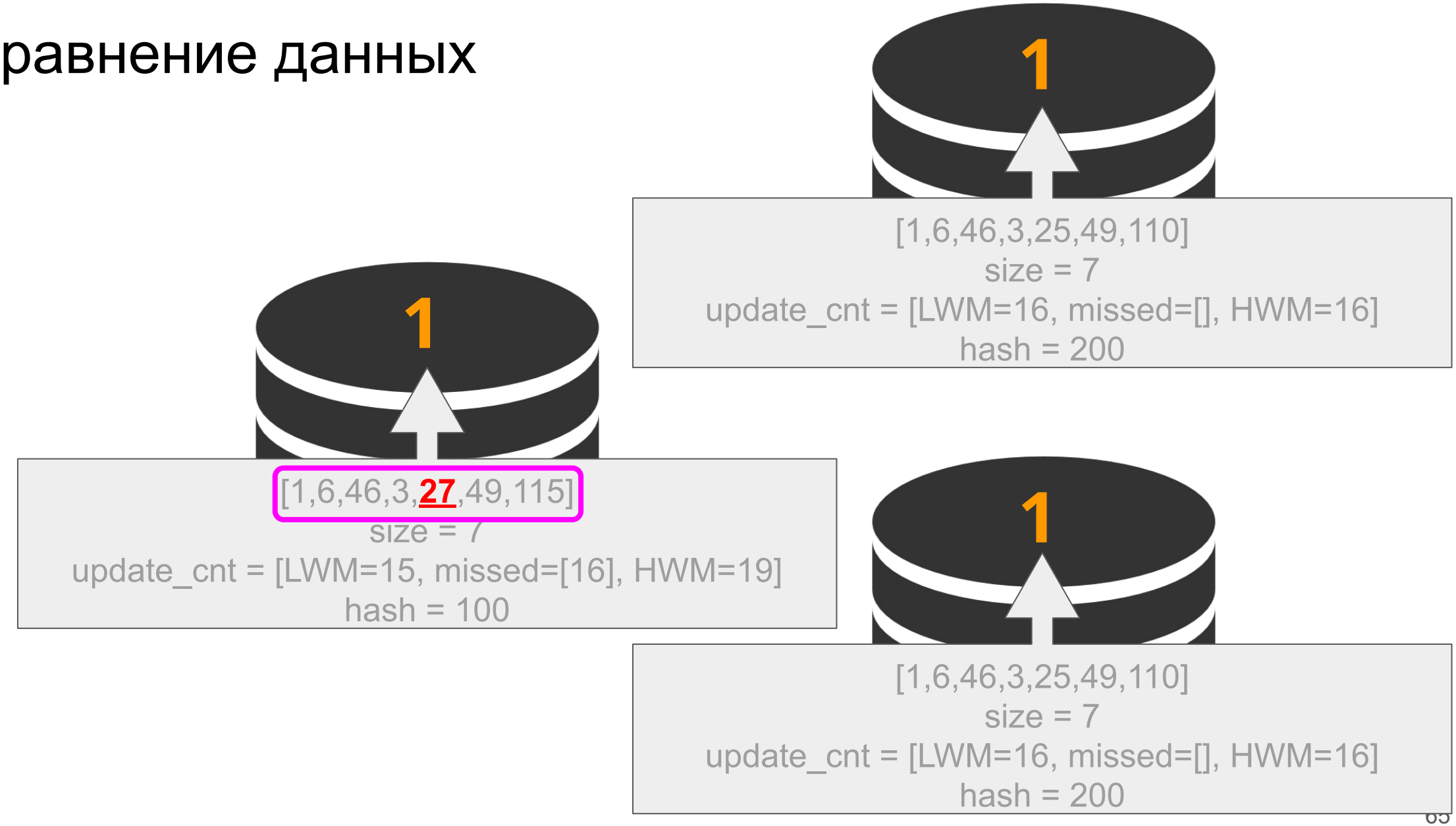
#5 Сравнение данных



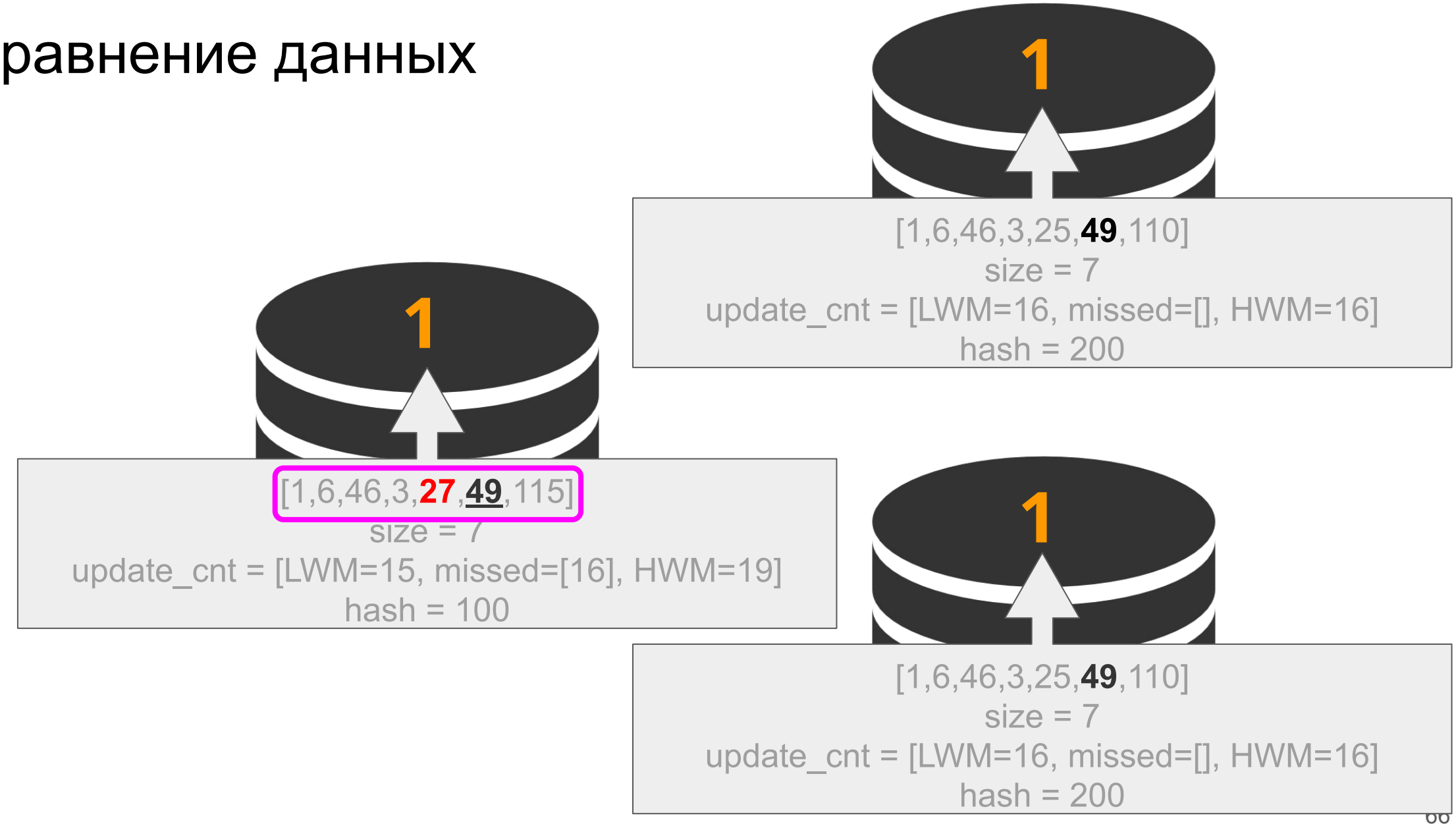
#5 Сравнение данных



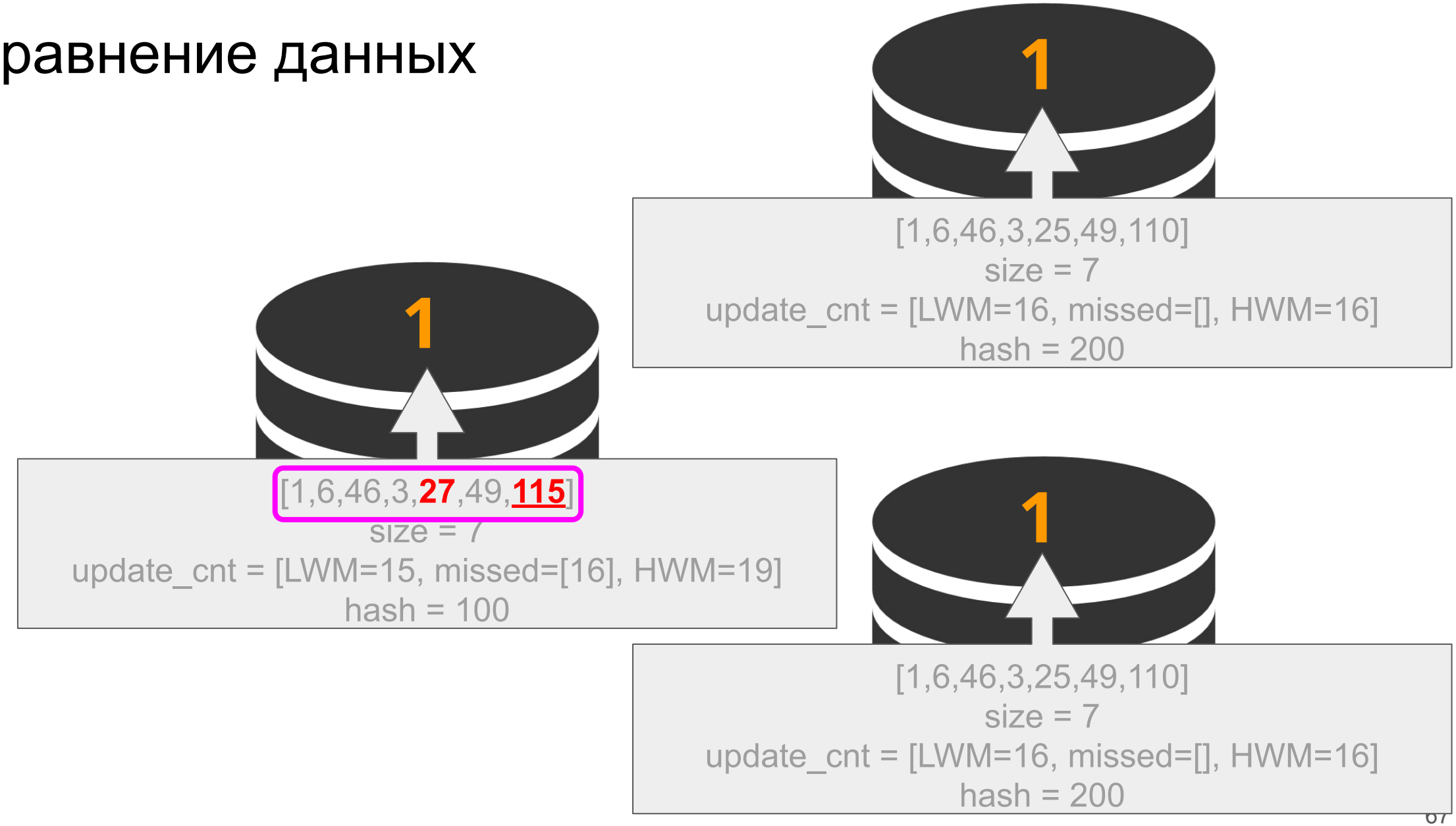
#5 Сравнение данных



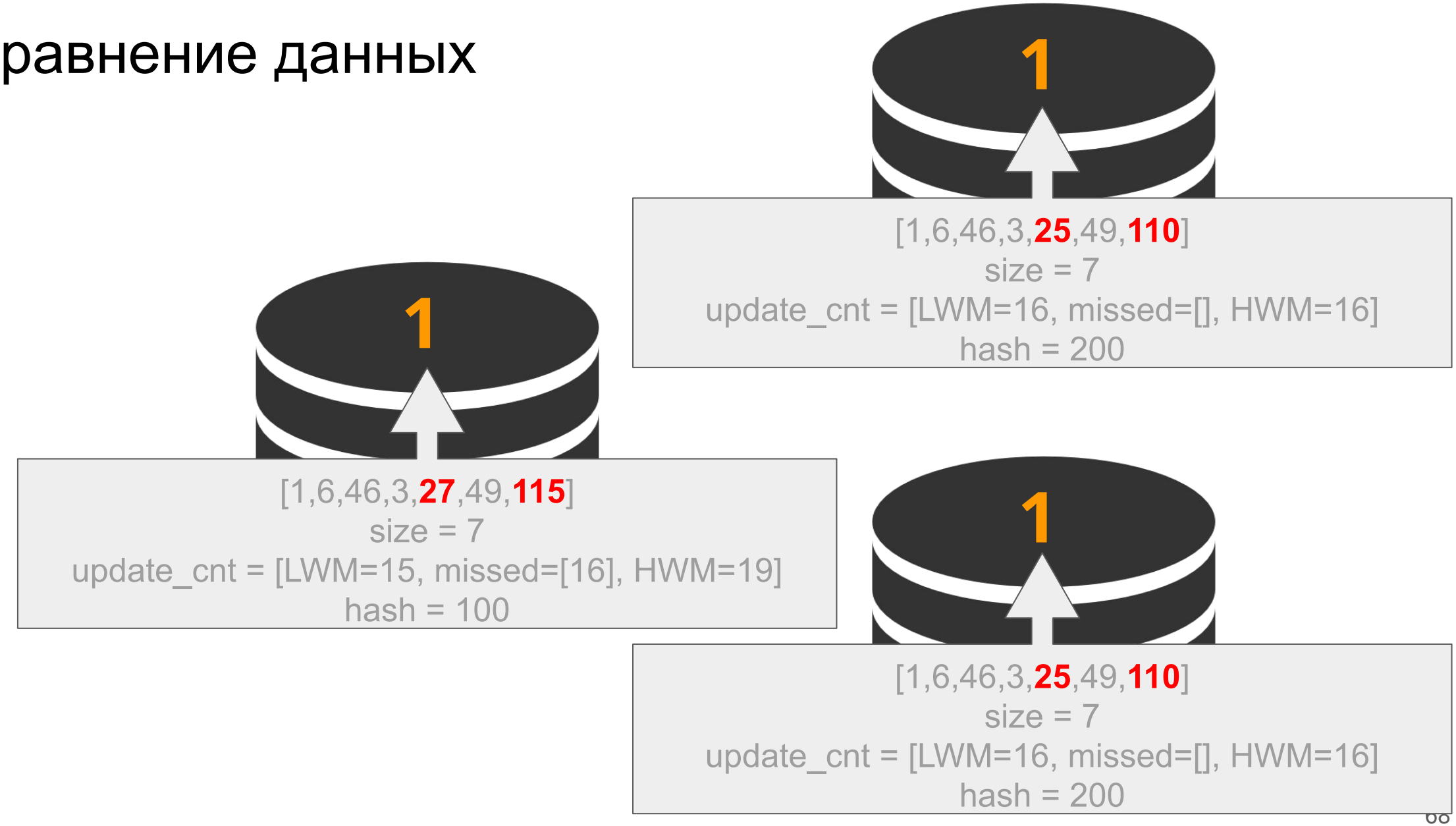
#5 Сравнение данных



#5 Сравнение данных



#5 Сравнение данных



Стратегии починки

- LWW
- PRIMARY
- RELATIVE_MAJORITY
- REMOVE
- CHECK_ONLY

Стратегии починки

- LWW
- PRIMARY
- RELATIVE_MAJORITY
- REMOVE
- **CHECK_ONLY**

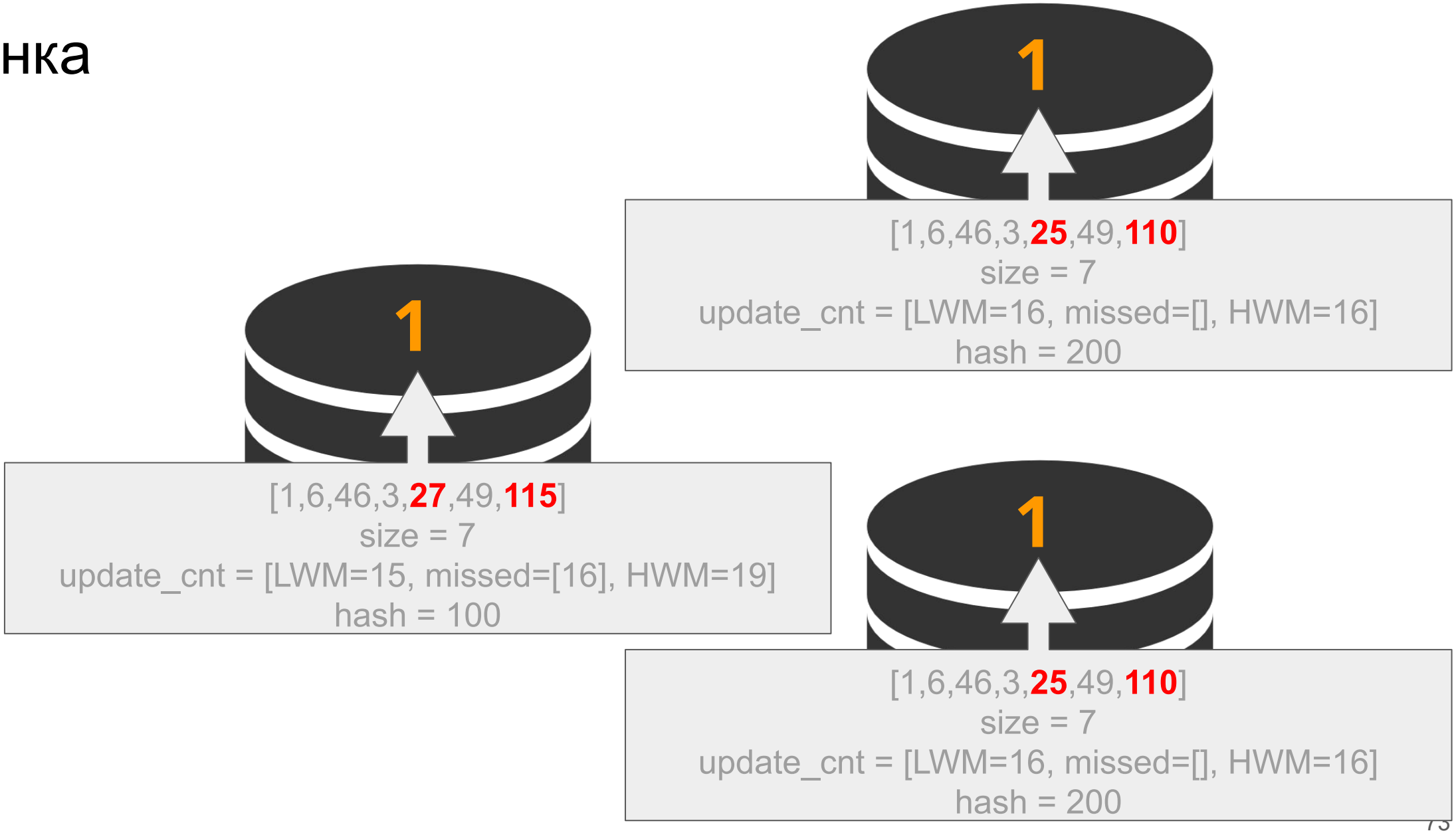
Стратегии починки

- **LWW**
- PRIMARY
- **RELATIVE_MAJORITY**
- REMOVE
- CHECK_ONLY

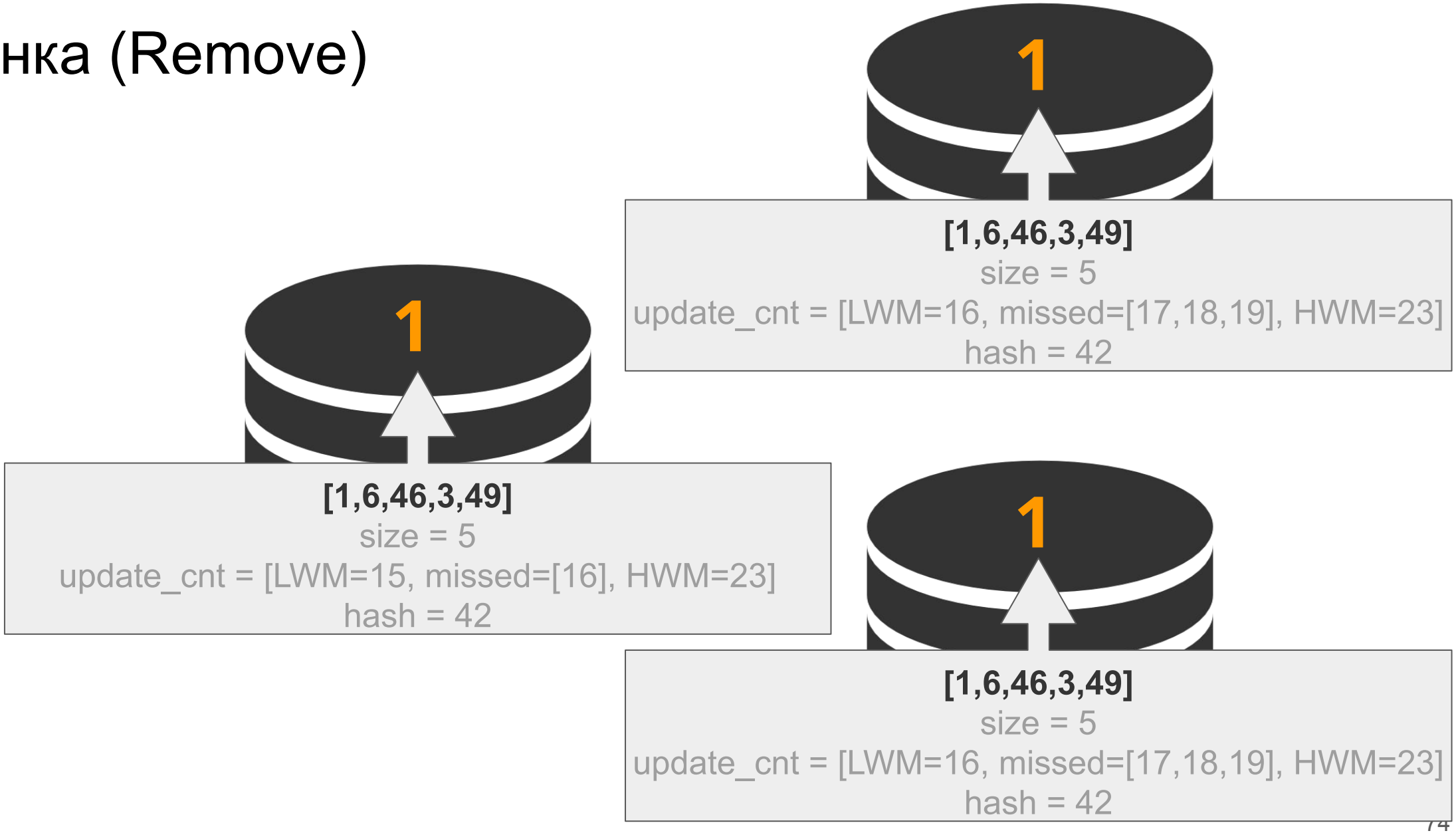
Стратегии починки

- LWW
- **PRIMARY**
- RELATIVE_MAJORITY
- **REMOVE**
- CHECK_ONLY

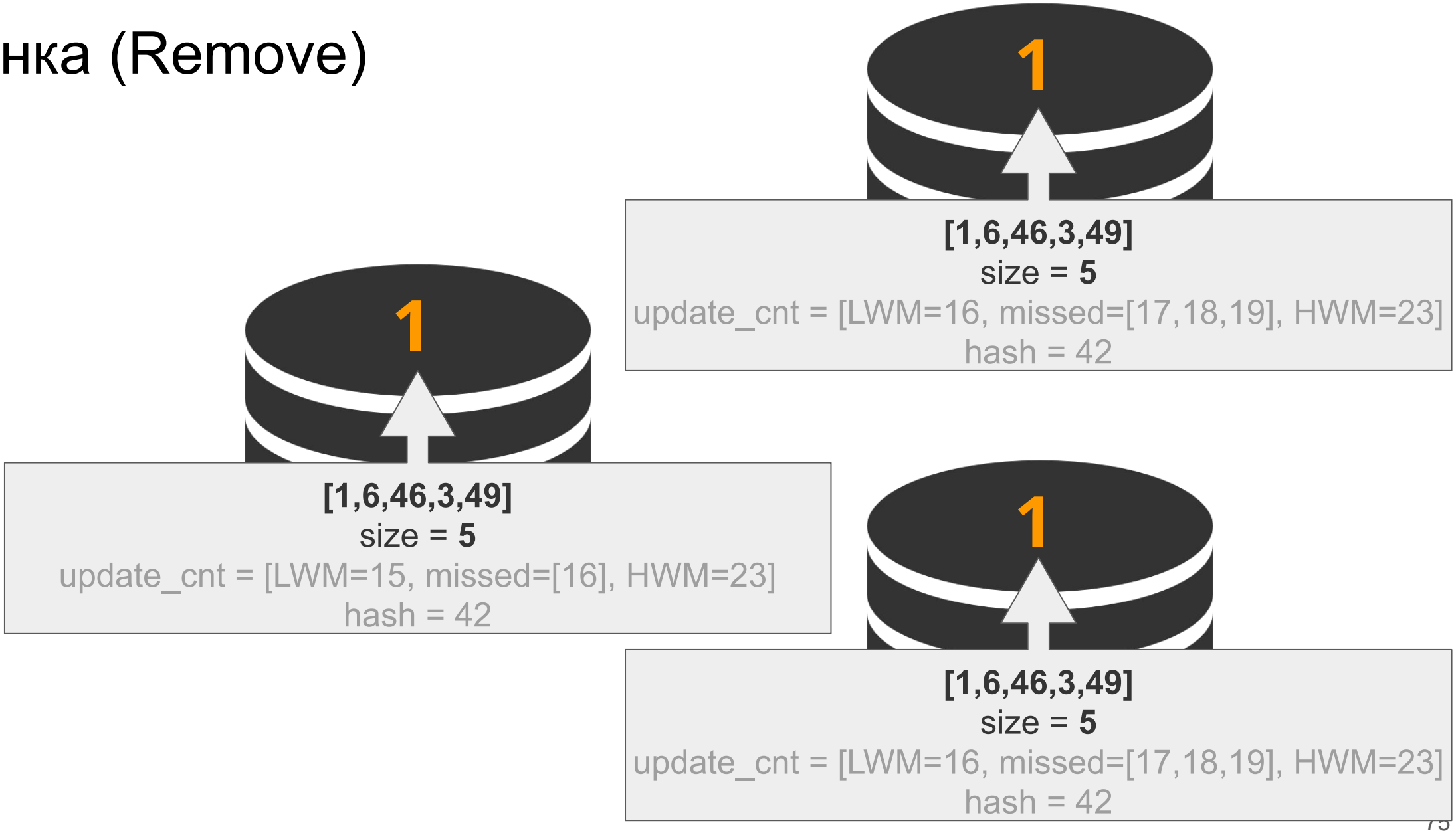
Починка


1
[1,6,46,3,25,49,110]
size = 7
update_cnt = [LWM=16, missed=[], HWM=16]
hash = 200
1
[1,6,46,3,27,49,115]
size = 7
update_cnt = [LWM=15, missed=[16], HWM=19]
hash = 100
1
[1,6,46,3,25,49,110]
size = 7
update_cnt = [LWM=16, missed=[], HWM=16]
hash = 200

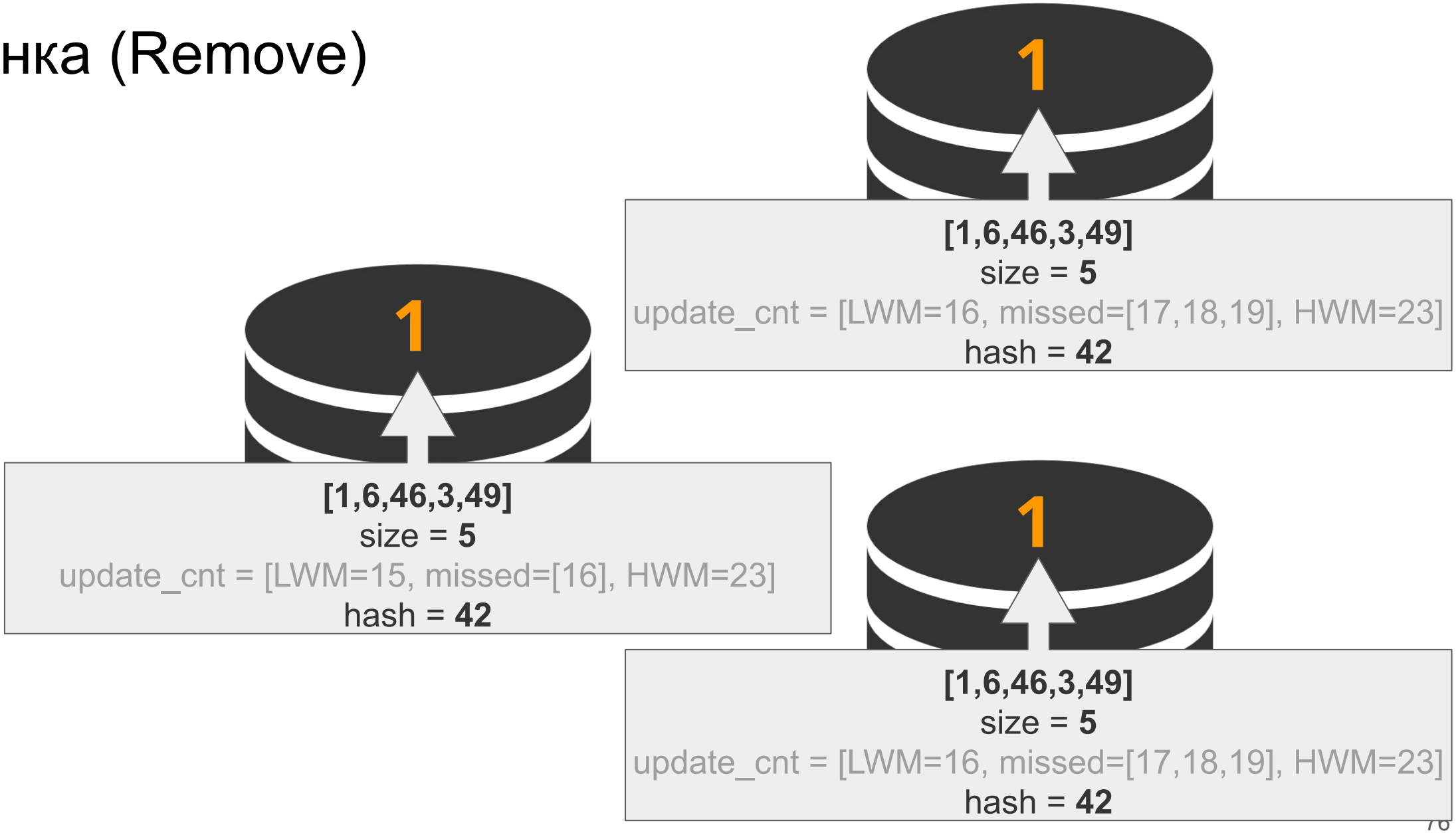
Починка (Remove)



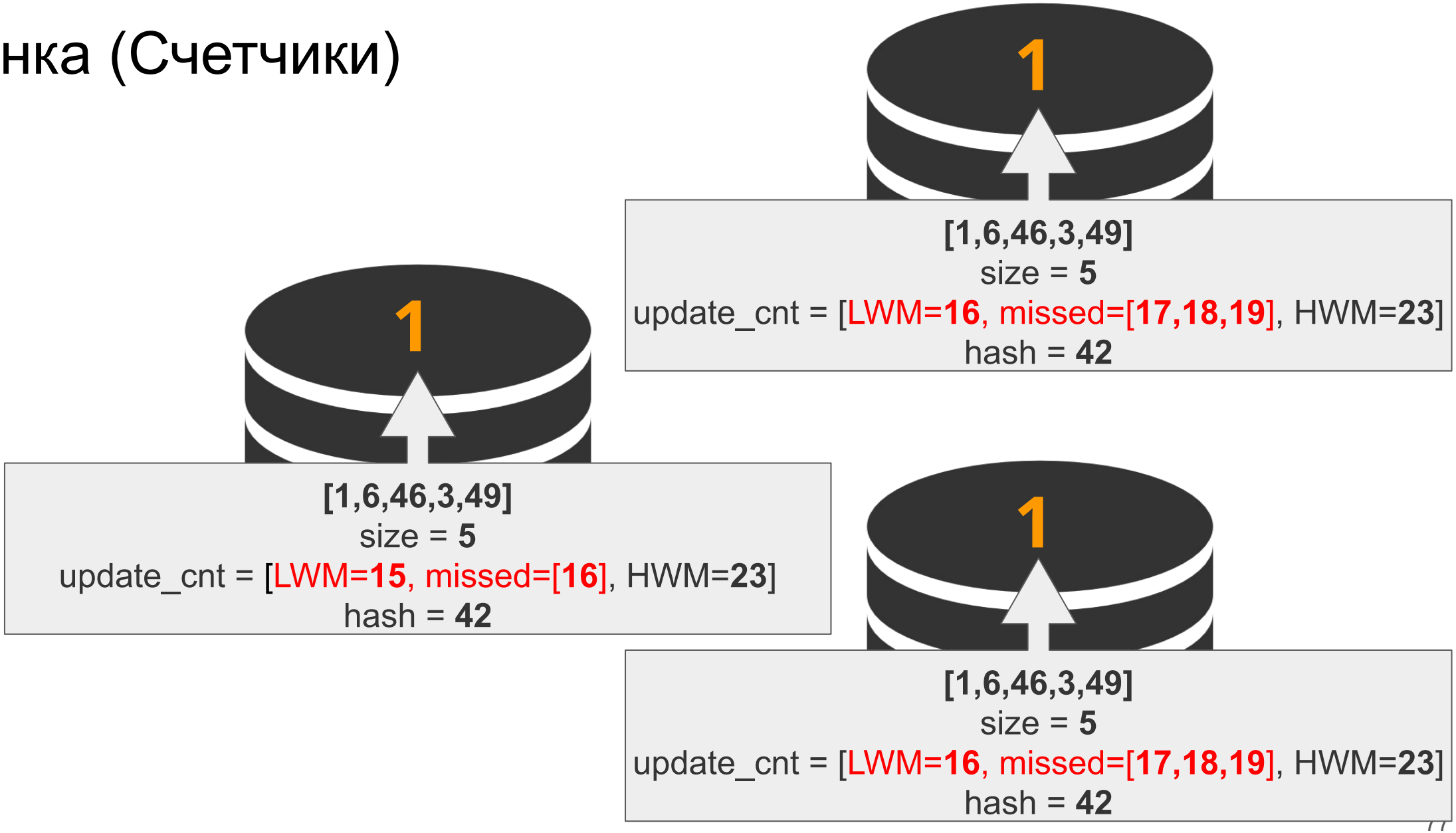
Починка (Remove)



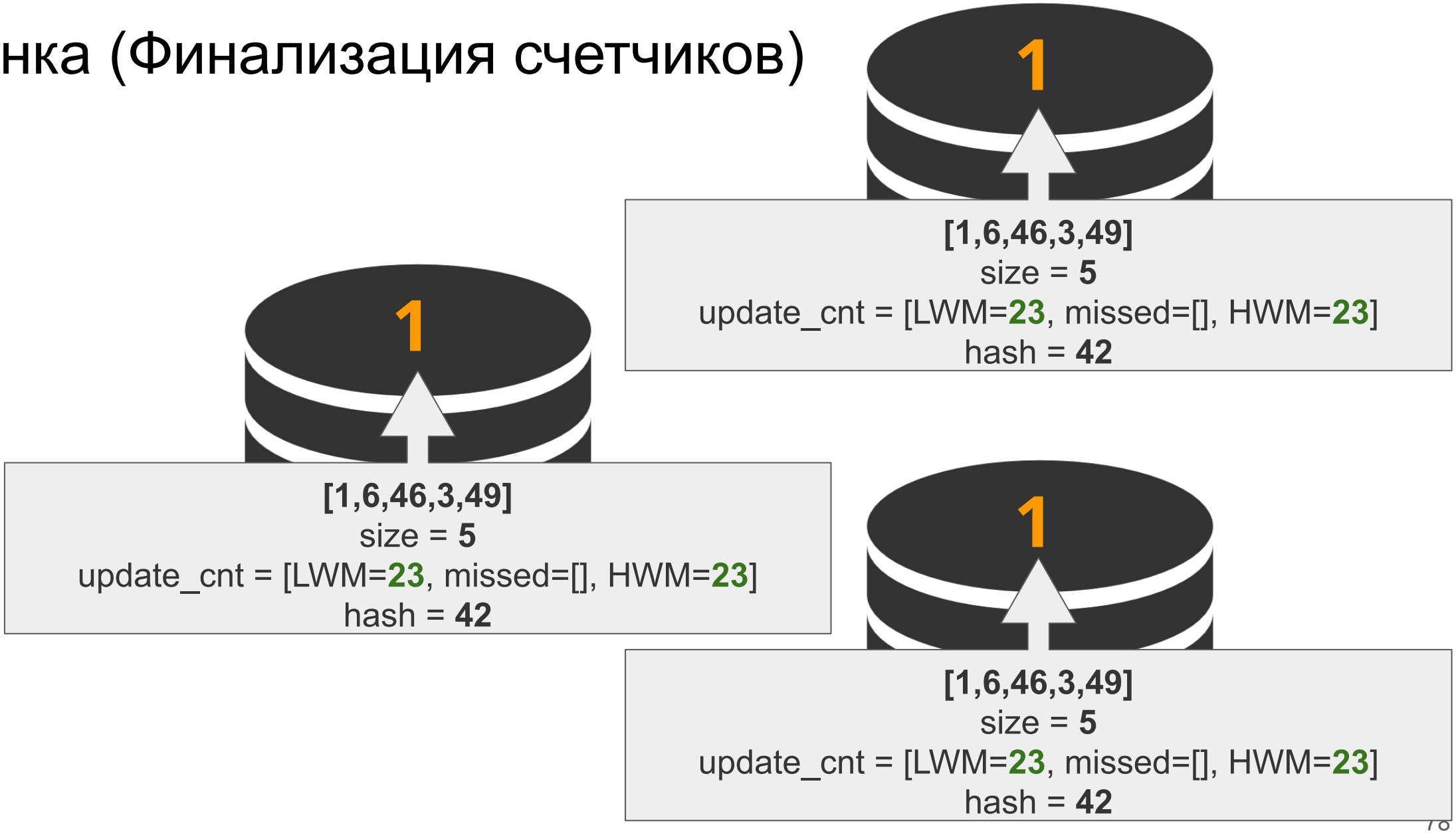
Починка (Remove)



Починка (Счетчики)



Починка (Финализация счетчиков)



Data consistency

https://cwiki.apache.org/confluence/display/IGNITE/Data+consistency

Read Repair

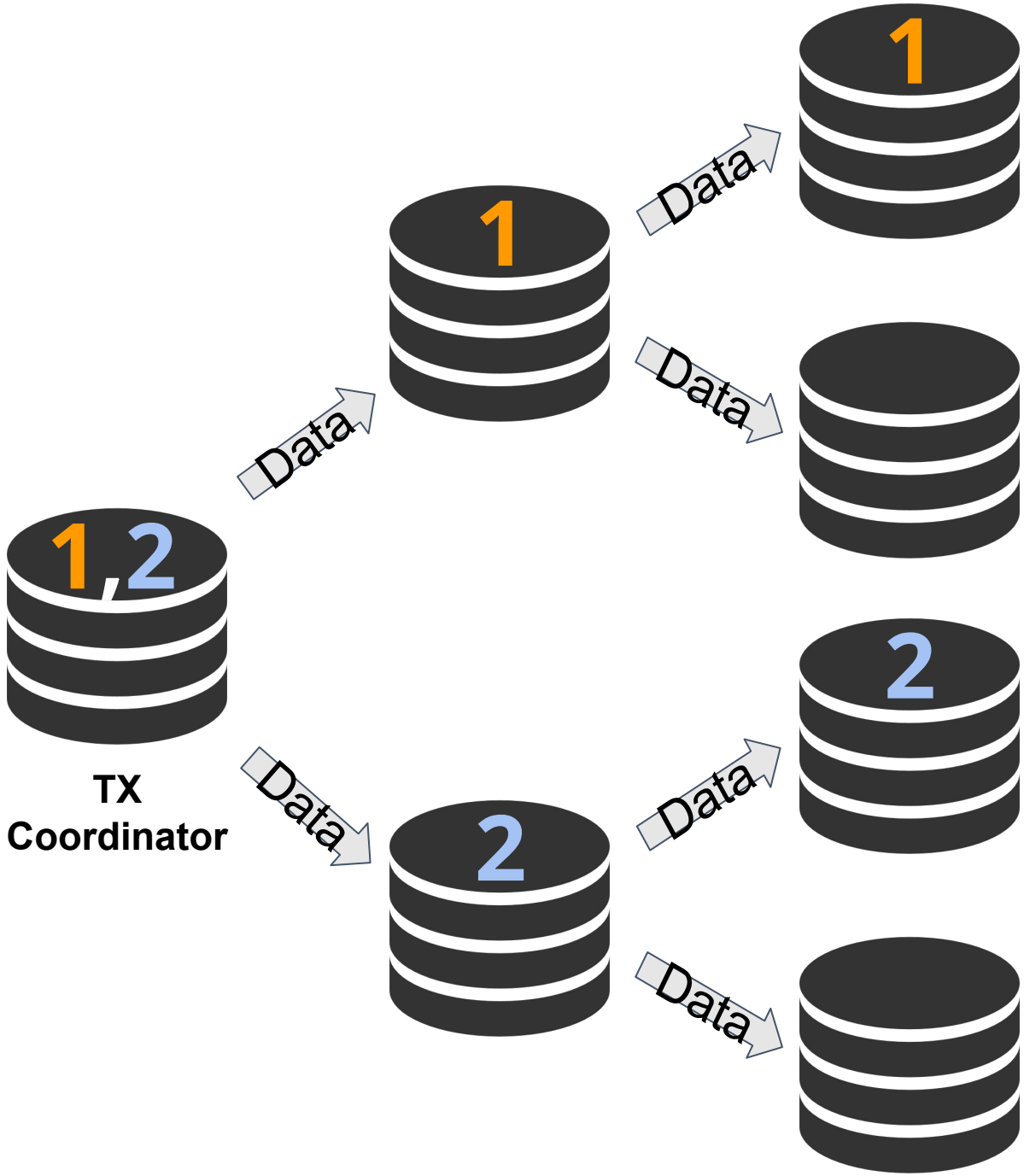
https://ignite.apache.org/docs/latest/key-value-api/read-repair

Consistency Check and Repair

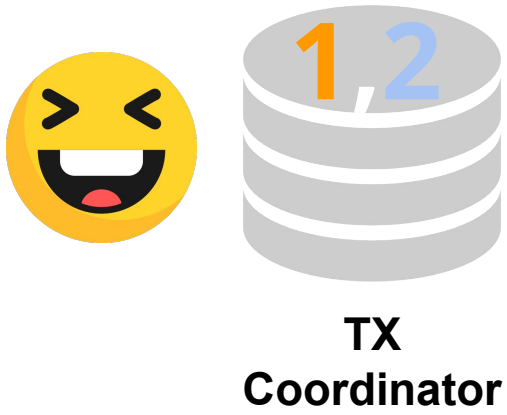
https://ignite.apache.org/docs/latest/tools/control-script#consistency-check-and-repair-commands


THIS IS FINE.

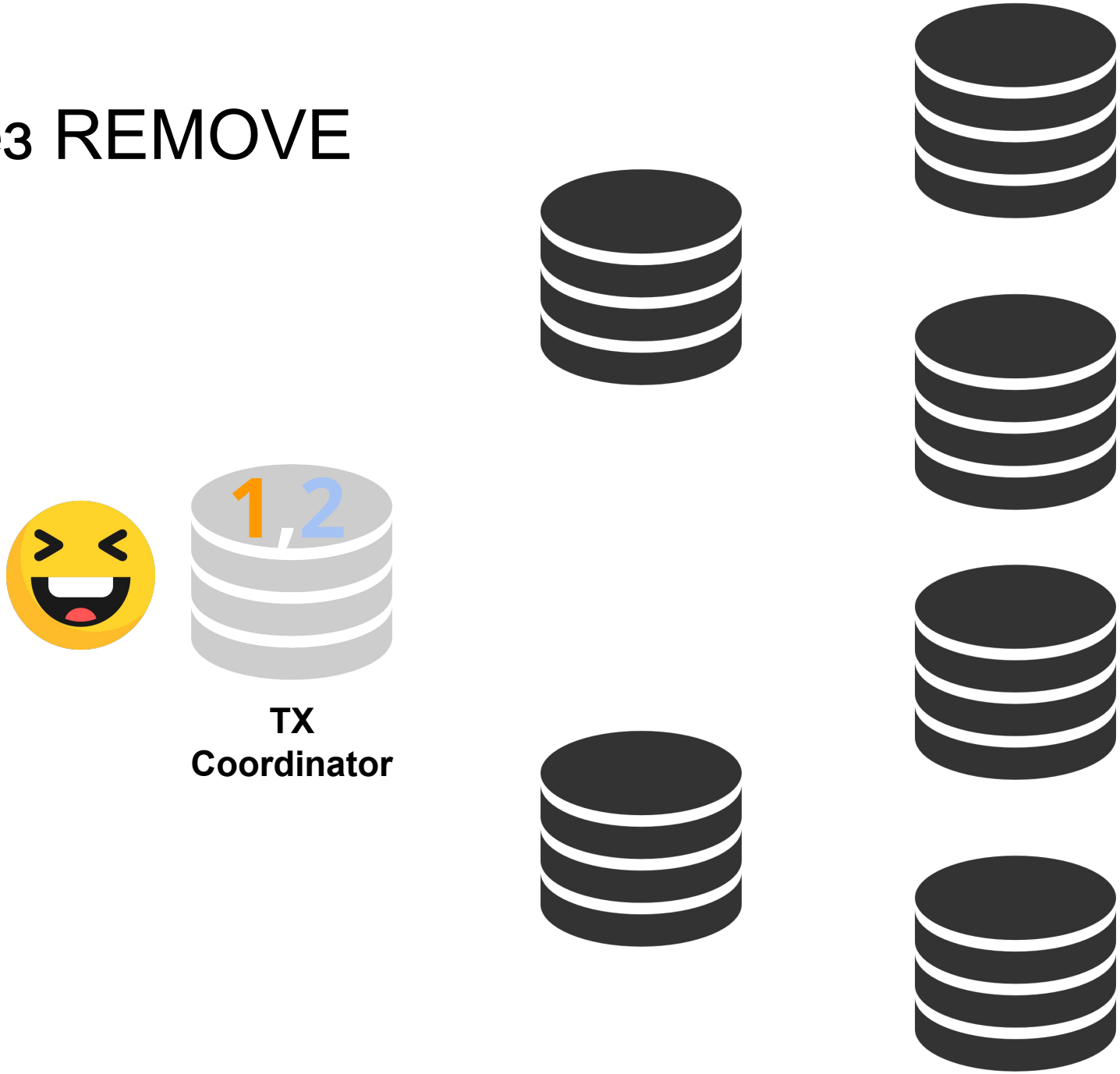
Распределенные транзакции



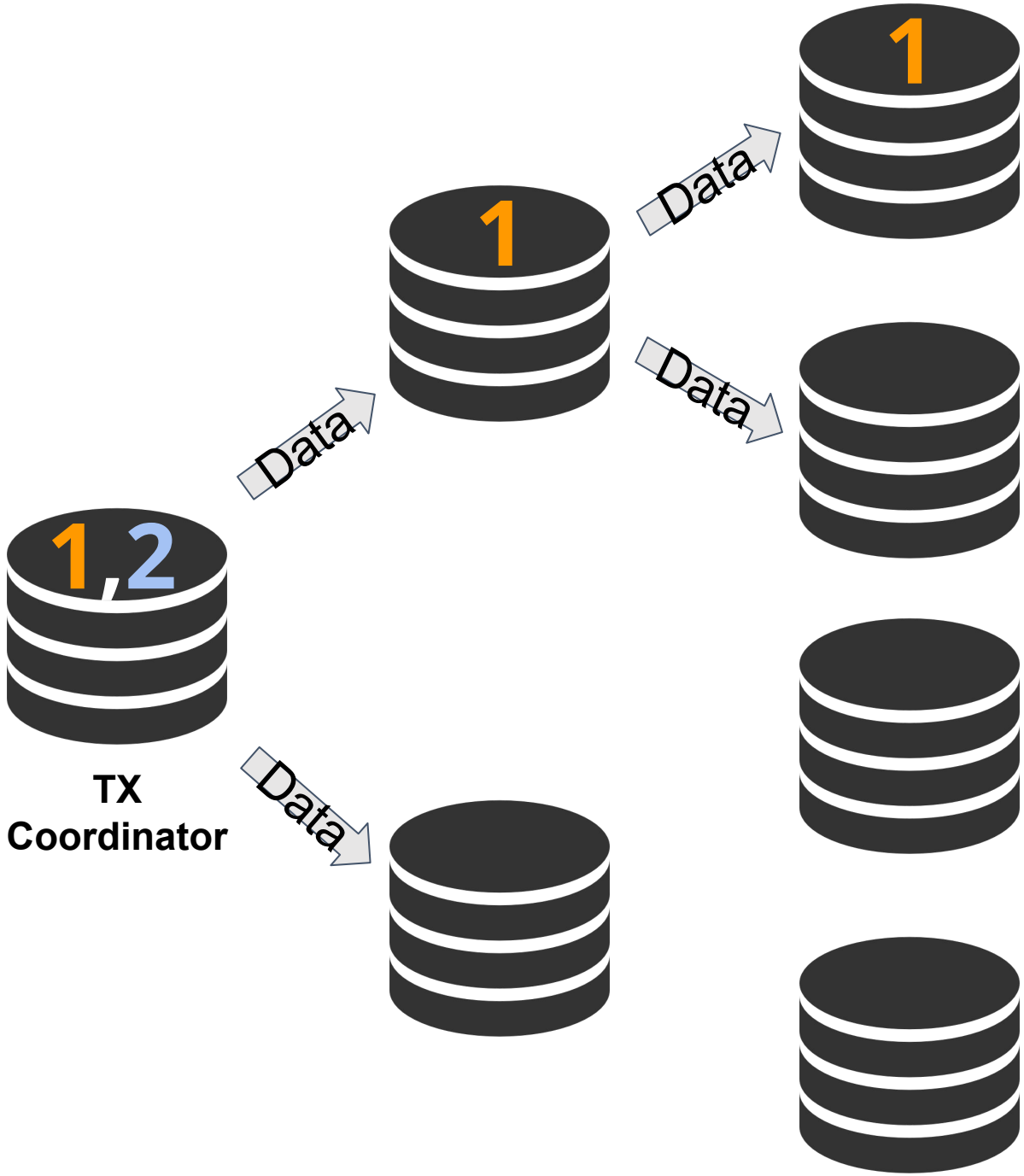
Чиним через PRIMARY



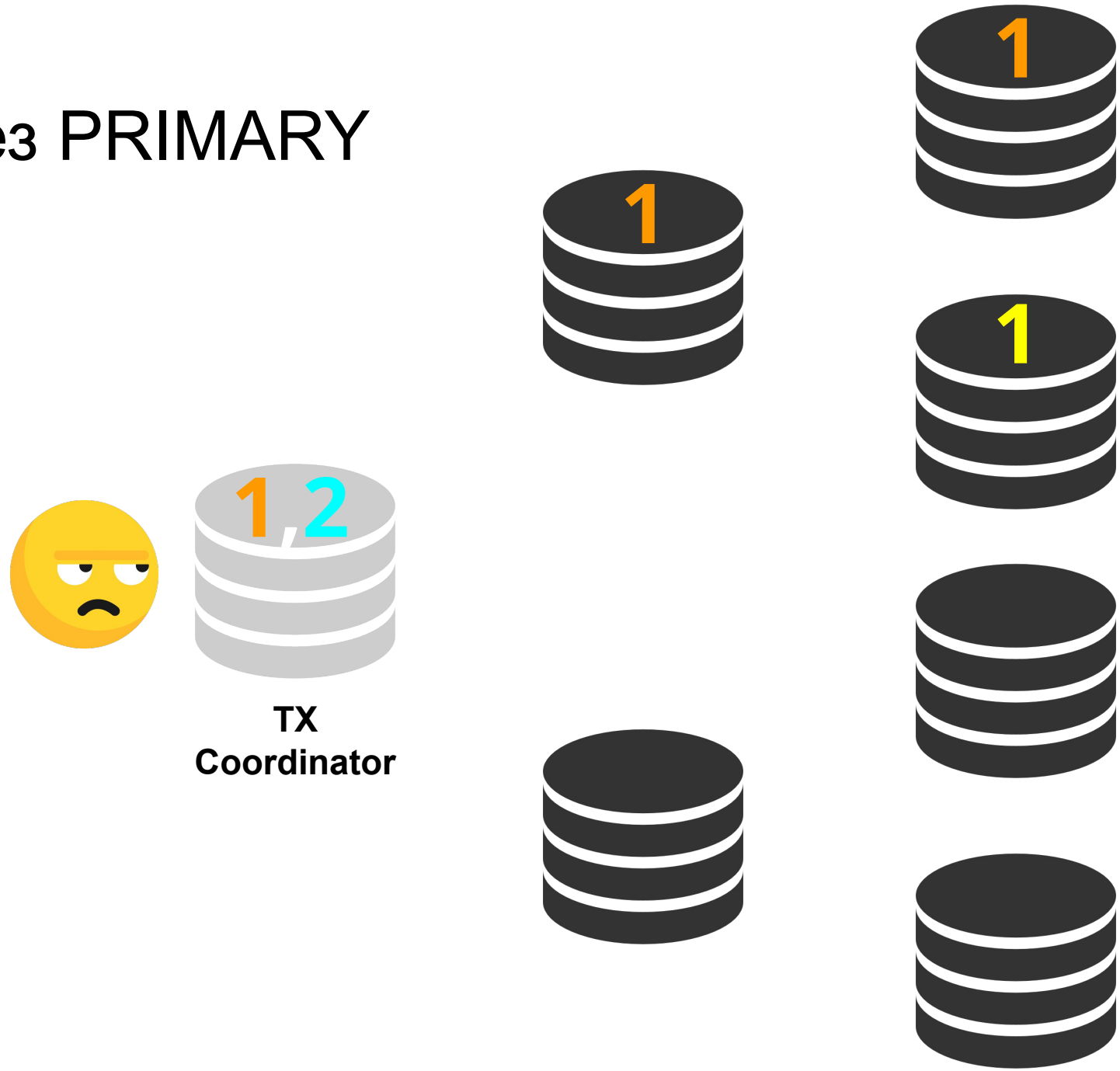
Чиним через REMOVE



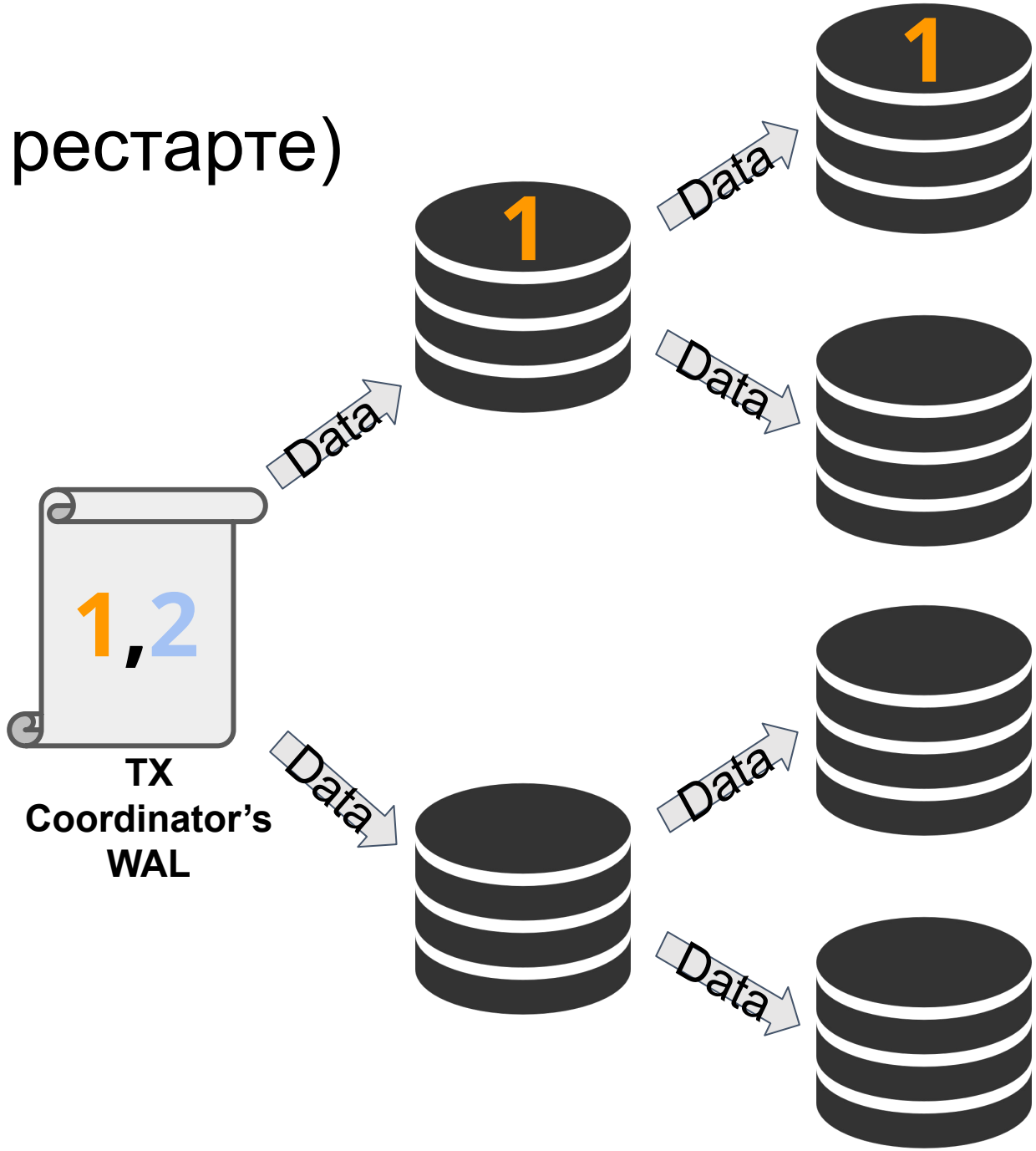
TX Consistency?



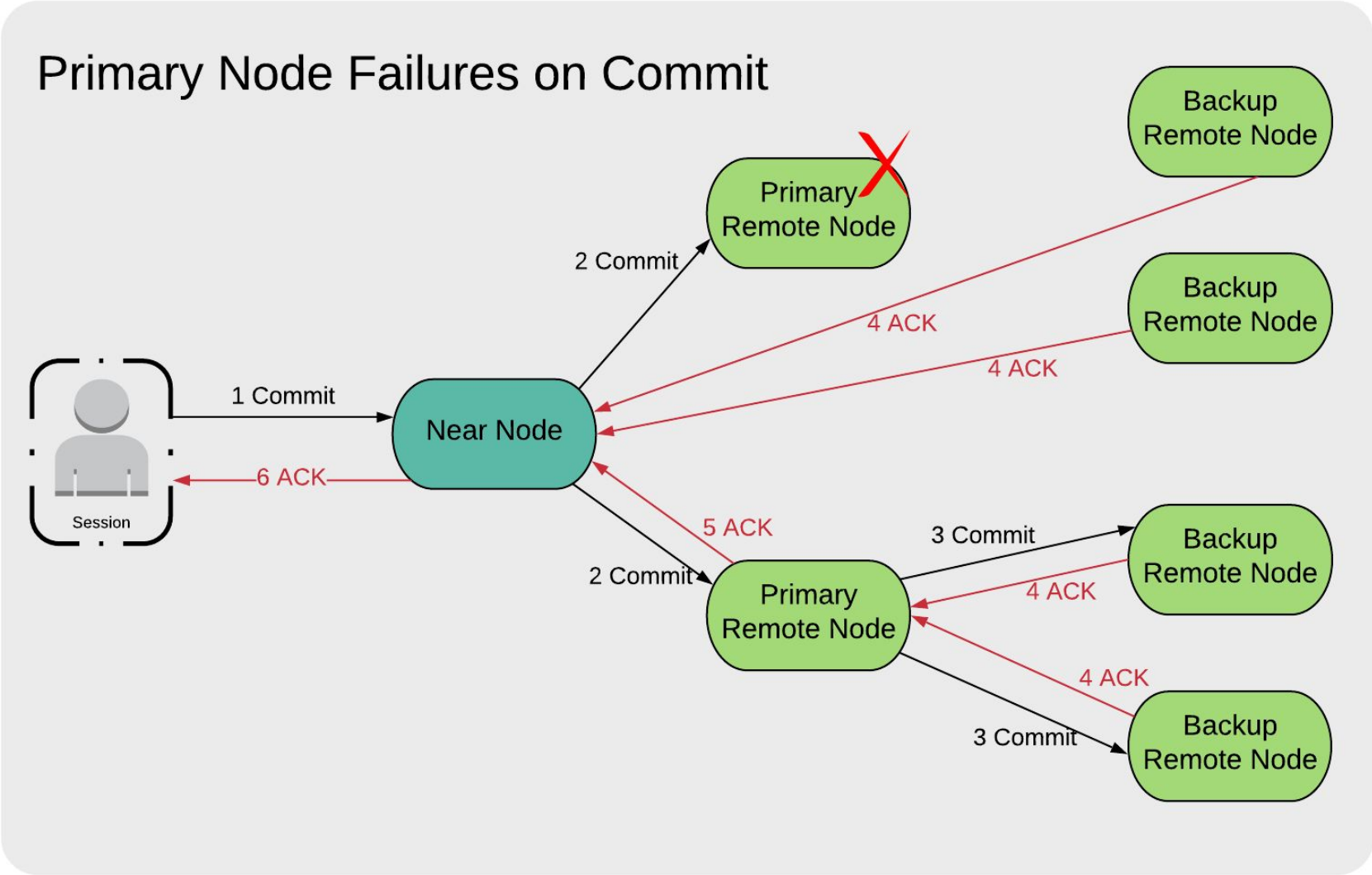
Чиним через PRIMARY



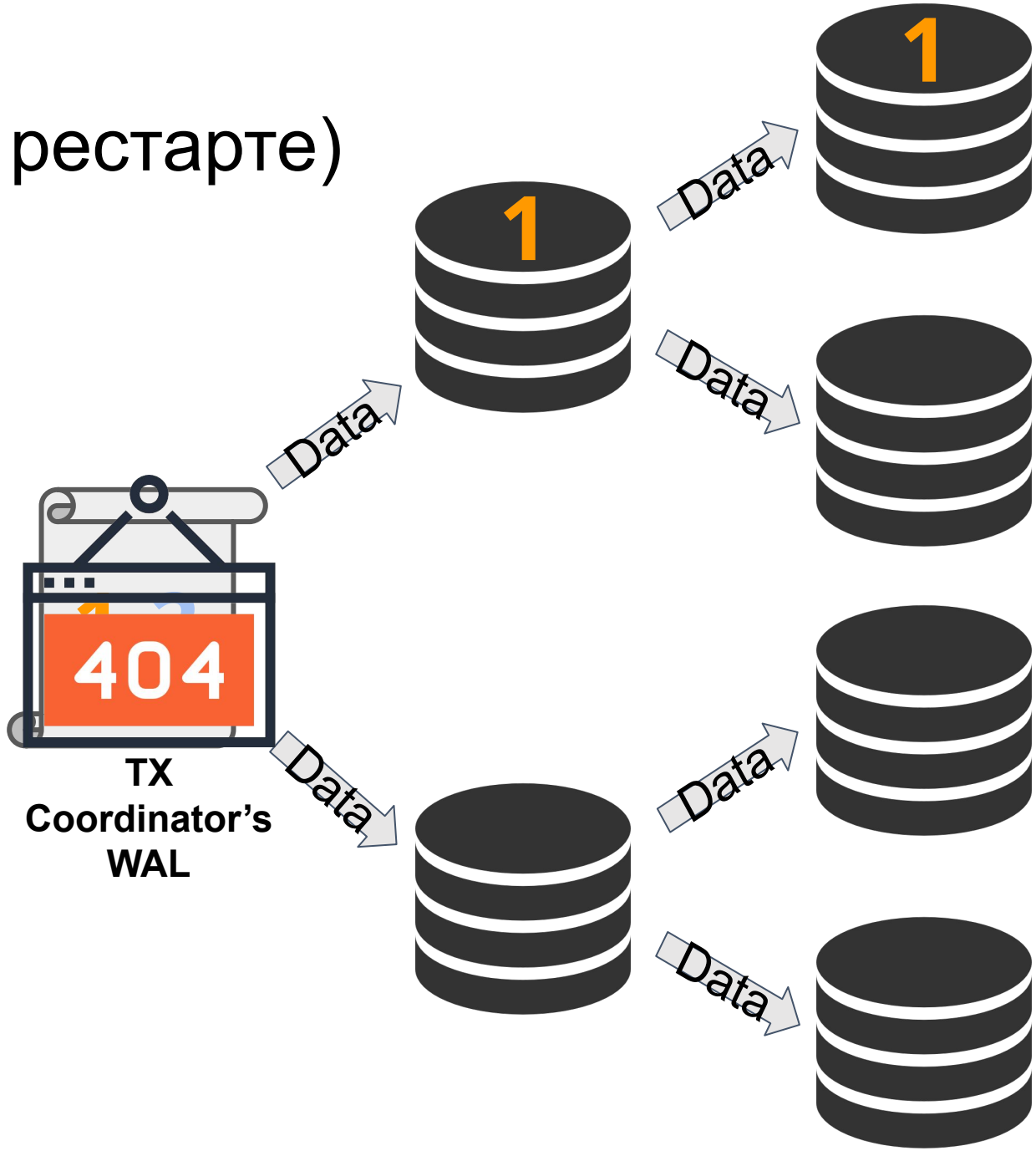
TX Recovery (на рестарте)



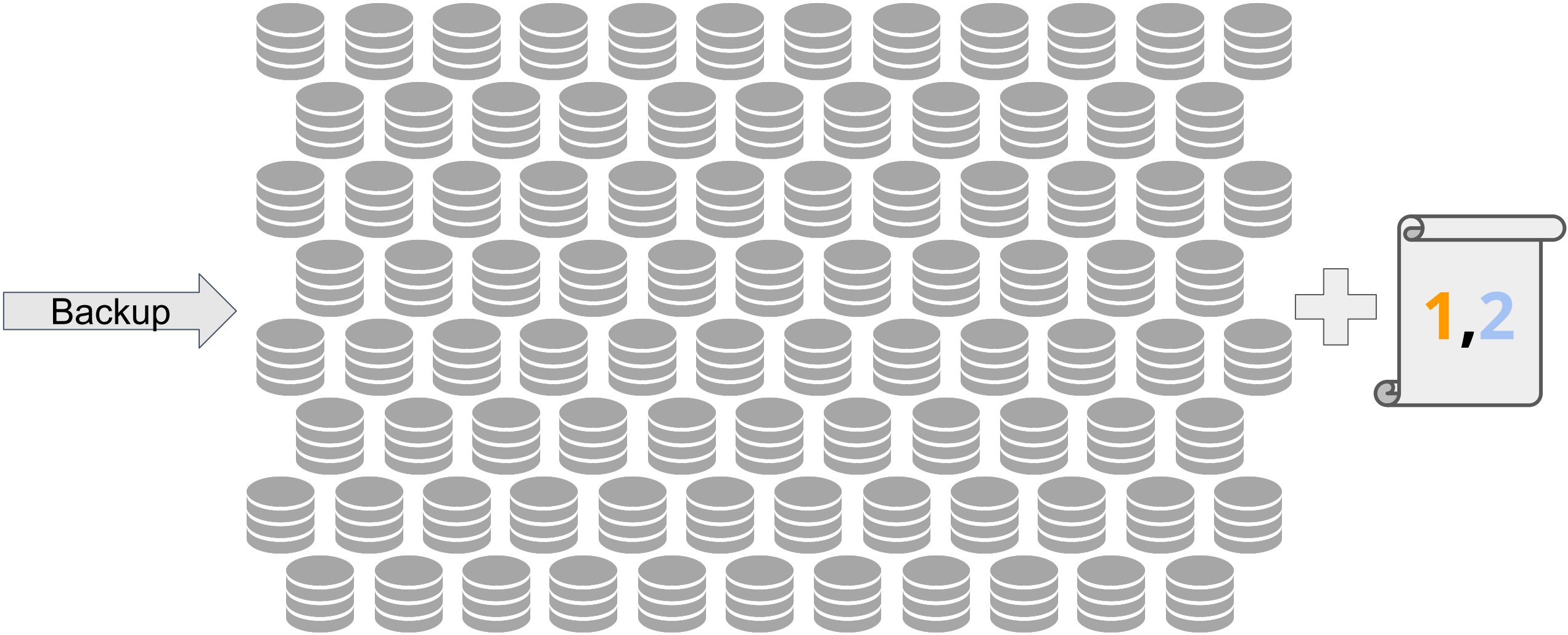
TX Recovery (на выход узла/узлов)



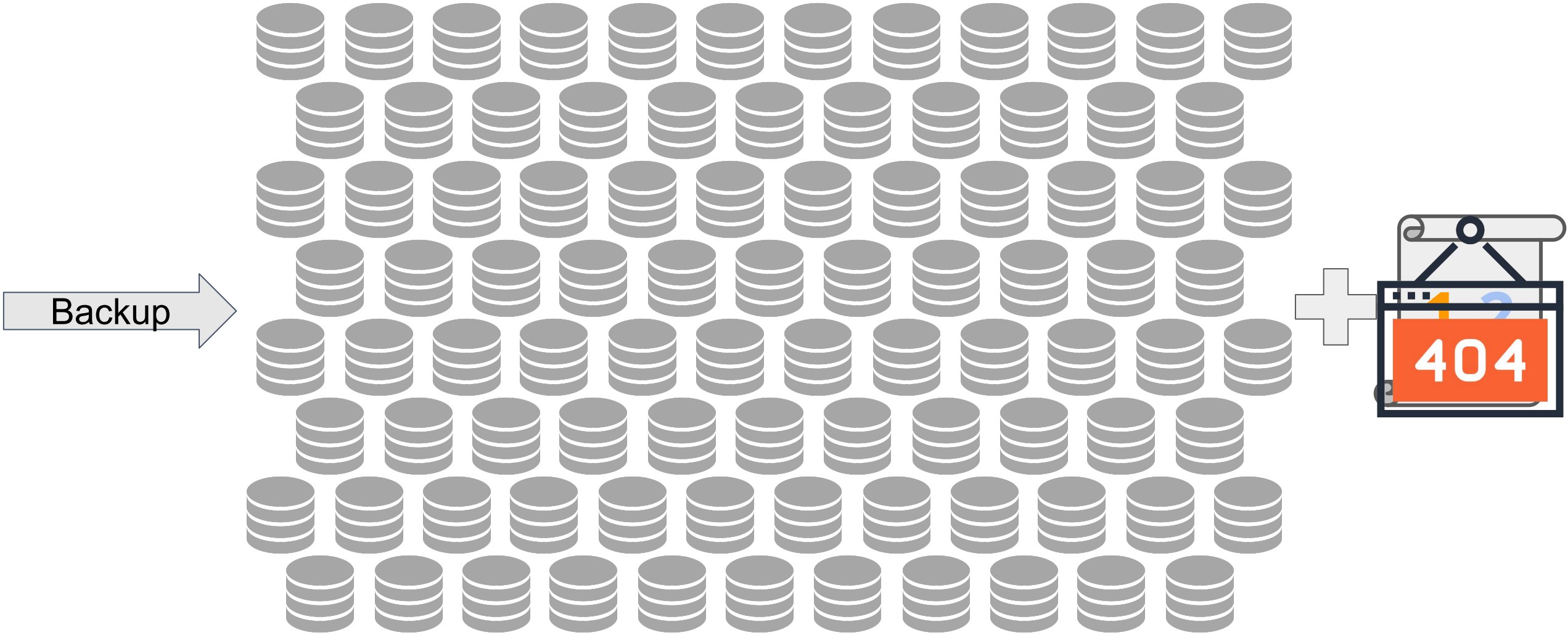
TX Recovery (на рестарте)



Бэкапы + WAL



Бэкапы + WAL



Checklist гарантий восстановления после аварии

- Автоматические фэйловеры
- Ручные инструменты

Checklist гарантий восстановления после аварии

- Автоматические фэйловеры
- Ручные инструменты
- **Избегать аварии**

Checklist гарантий восстановления после аварии

- Автоматические фэйловеры
- Ручные инструменты
- **Избегать аварии**