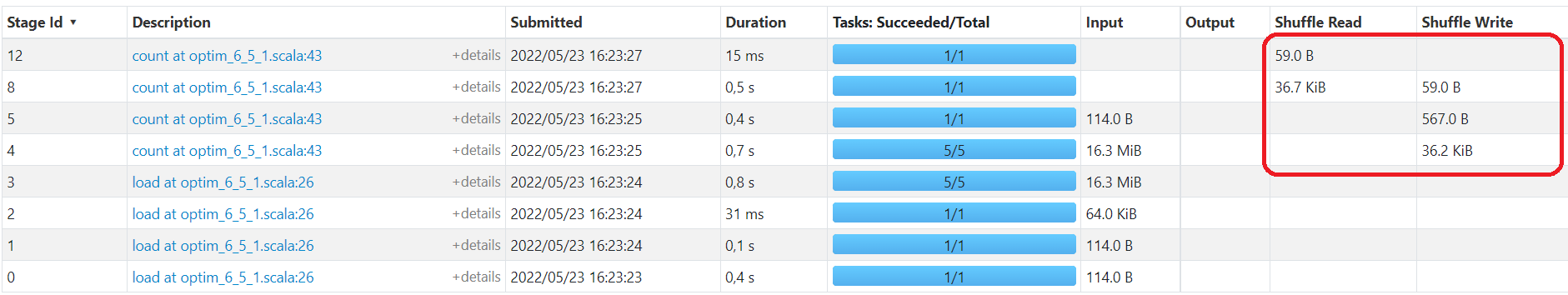
**ОТВЕТ - 42996**

**Анализ неоптимального выполнения**

Файл hotel\_bookings.csv содержит 119390 строк.

Файл hotel\_ cancel\_codes.csv - всего 8 строк.

При попытке соединить эти 2 файла требуется со всех разделов собрать, чтобы строки друг с другом соединить по условию соединения, и заново распределить из одного датафрейма (*hotelsDF*) 119390 строк, а из другого (*codesDF)* всего 8 строк. В статистике Spark UI это перераспределение можно увидеть:



В плане выполнения видим эту перетасовку (Exchange):

== Physical Plan ==

AdaptiveSparkPlan isFinalPlan=false

+- SortMergeJoin [is\_canceled#39], [id#16], Inner, NOT ((reservation\_status#68 = Canceled) = (is\_cancelled#17 = yes))

:- Sort [is\_canceled#39 ASC NULLS FIRST], false, 0

: +- **Exchange hashpartitioning(is\_canceled#39, 200), ENSURE\_REQUIREMENTS, [id=#60]**

: +- Filter (isnotnull(reservation\_status#68) AND isnotnull(is\_canceled#39))

: +- FileScan csv [hotel#38,is\_canceled#39,lead\_time#40,arrival\_date\_year#41,arrival\_date\_month#42,arrival\_date\_week\_number#43,arrival\_date\_day\_of\_month#44,stays\_in\_weekend\_nights#45,stays\_in\_week\_nights#46,adults#47,children#48,babies#49,meal#50,country#51,market\_segment#52,distribution\_channel#53,is\_repeated\_guest#54,previous\_cancellations#55,previous\_bookings\_not\_canceled#56,reserved\_room\_type#57,assigned\_room\_type#58,booking\_changes#59,deposit\_type#60,agent#61,... 8 more fields] Batched: false, DataFilters: [isnotnull(reservation\_status#68), isnotnull(is\_canceled#39)], Format: CSV, Location: InMemoryFileIndex(1 paths)[file:/C:/Scala/spark3/src/main/resources/hotel\_bookings.csv], PartitionFilters: [], PushedFilters: [IsNotNull(reservation\_status), IsNotNull(is\_canceled)], ReadSchema: struct<hotel:string,is\_canceled:int,lead\_time:int,arrival\_date\_year:int,arrival\_date\_month:string...

+- Sort [id#16 ASC NULLS FIRST], false, 0

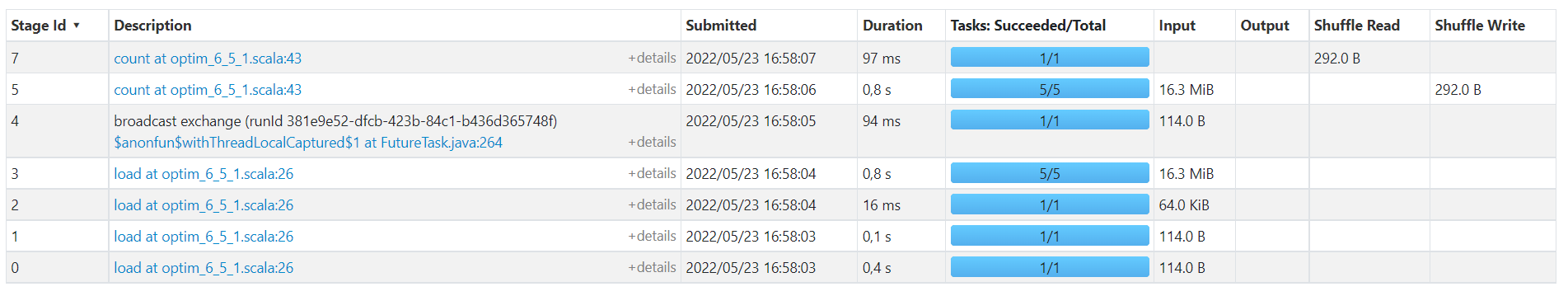
+- **Exchange hashpartitioning(id#16, 200), ENSURE\_REQUIREMENTS, [id=#61]**

+- Filter (isnotnull(is\_cancelled#17) AND isnotnull(id#16))

+- FileScan csv [id#16,is\_cancelled#17,confirmation#18] Batched: false, DataFilters: [isnotnull(is\_cancelled#17), isnotnull(id#16)], Format: CSV, Location: InMemoryFileIndex(1 paths)[file:/C:/Scala/spark3/src/main/resources/cancel\_codes.csv], PartitionFilters: [], PushedFilters: [IsNotNull(is\_cancelled), IsNotNull(id)], ReadSchema: struct<id:int,is\_cancelled:string,confirmation:string>

**Анализ оптимального выполнения**

Оптимизировать будем функцией broadcast. Применив её к датафрейму с кодами отмены *codesDF*, этот датафрейм будет скопирован на каждый из executor к частям датафрейма с данными по отелям *hotelsDF*. При её применении количество перераспределенных данных резко уменьшилось – не потребовалось перемещать строки из *hotelsDF*.



И улучшился также и план выполнения. В нём нет Exchange:

== Physical Plan ==

AdaptiveSparkPlan isFinalPlan=false

+- BroadcastHashJoin [is\_canceled#39], [id#16], Inner, BuildRight, NOT ((reservation\_status#68 = Canceled) = (is\_cancelled#17 = yes)), false

:- Filter (isnotnull(reservation\_status#68) AND isnotnull(is\_canceled#39))

: +- FileScan csv [hotel#38,is\_canceled#39,lead\_time#40,arrival\_date\_year#41,arrival\_date\_month#42,arrival\_date\_week\_number#43,arrival\_date\_day\_of\_month#44,stays\_in\_weekend\_nights#45,stays\_in\_week\_nights#46,adults#47,children#48,babies#49,meal#50,country#51,market\_segment#52,distribution\_channel#53,is\_repeated\_guest#54,previous\_cancellations#55,previous\_bookings\_not\_canceled#56,reserved\_room\_type#57,assigned\_room\_type#58,booking\_changes#59,deposit\_type#60,agent#61,... 8 more fields] Batched: false, DataFilters: [isnotnull(reservation\_status#68), isnotnull(is\_canceled#39)], Format: CSV, Location: InMemoryFileIndex(1 paths)[file:/C:/Scala/spark3/src/main/resources/hotel\_bookings.csv], PartitionFilters: [], PushedFilters: [IsNotNull(reservation\_status), IsNotNull(is\_canceled)], ReadSchema: struct<hotel:string,is\_canceled:int,lead\_time:int,arrival\_date\_year:int,arrival\_date\_month:string...

+- BroadcastExchange HashedRelationBroadcastMode(List(cast(input[0, int, false] as bigint)),false), [id=#59]

+- Filter (isnotnull(is\_cancelled#17) AND isnotnull(id#16))

+- FileScan csv [id#16,is\_cancelled#17,confirmation#18] Batched: false, DataFilters: [isnotnull(is\_cancelled#17), isnotnull(id#16)], Format: CSV, Location: InMemoryFileIndex(1 paths)[file:/C:/Scala/spark3/src/main/resources/cancel\_codes.csv], PartitionFilters: [], PushedFilters: [IsNotNull(is\_cancelled), IsNotNull(id)], ReadSchema: struct<id:int,is\_cancelled:string,confirmation:string>