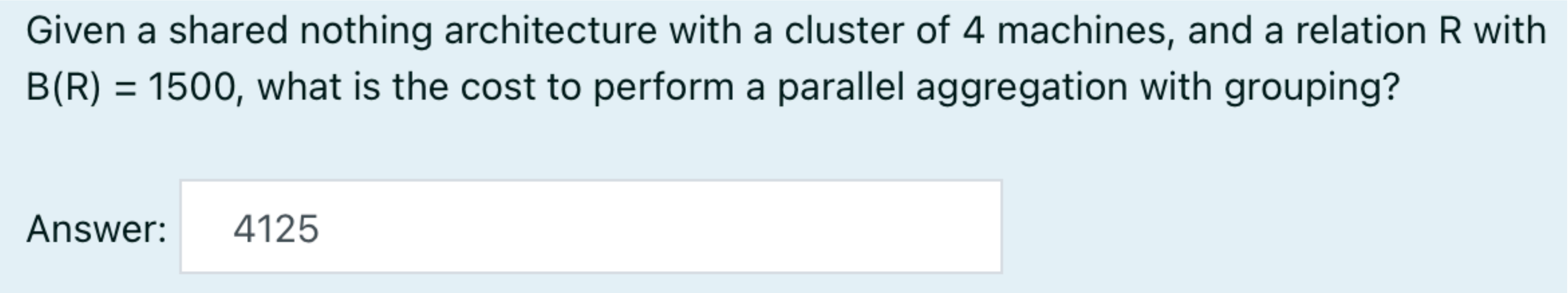
Quiz 10 Review

Parallel Databases and Big Data

Question 1

Parallel Sorting and Grouping Cost



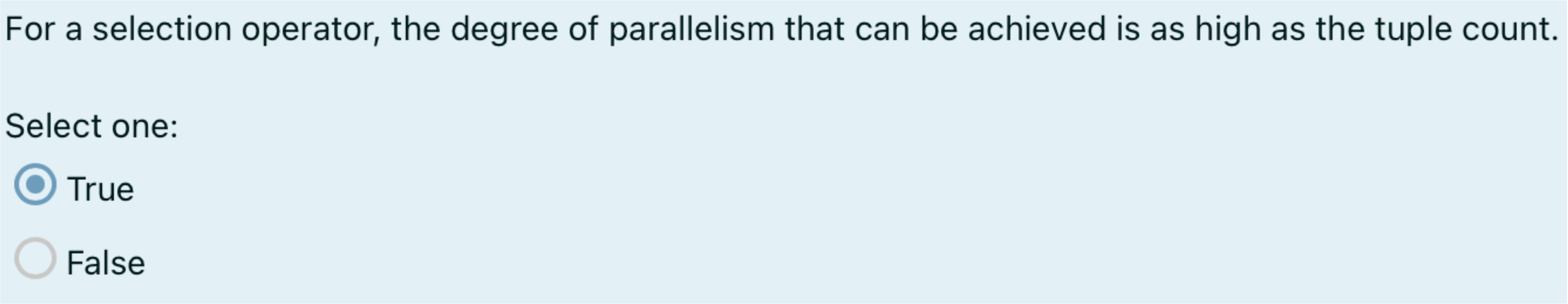
Given a shared nothing architecture with a cluster of 4 machines, and a relation R with B(R) = 1500, what is the cost to perform a parallel aggregation with grouping?

Answer: 4125

- Cost = B(R) partitioning + B(R) transfer + 3*B(R) / p local algorithm
- Cost = 1500 + 1500 + ((3*1500) / 4) = 3000 + 1125 = **4125**

Slide 35 in Parallel Databases

Question 4



For a selection operator, the degree of parallelism that can be achieved is as high as the tuple count.

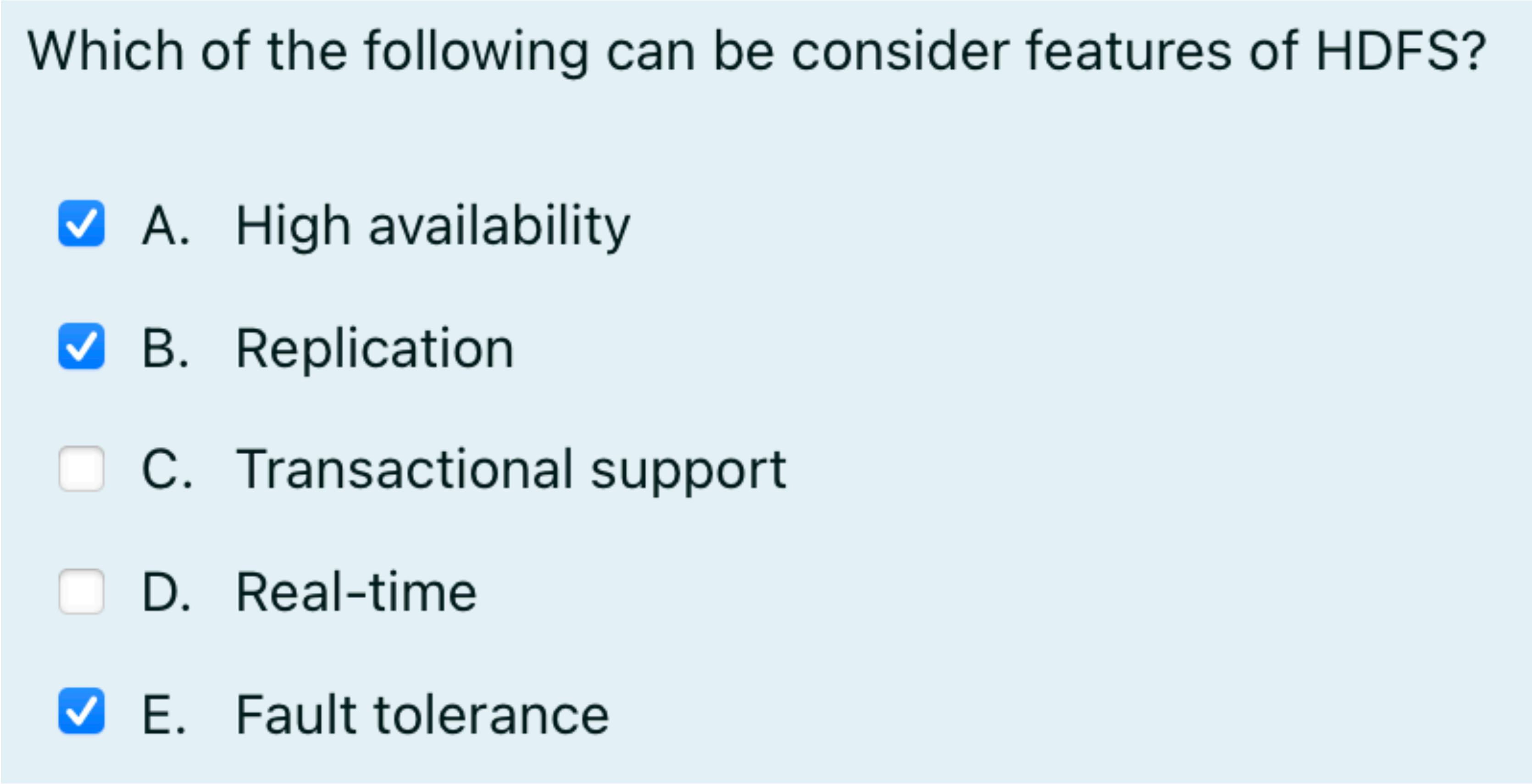
Select one:

● True

○ False

- This is because you can assign one selection-thread per tuple and run all of them in parallel. We can do so because the selection operator does not require any other information besides the input tuple, which makes each selection operator instance independent from each other.

Question 5

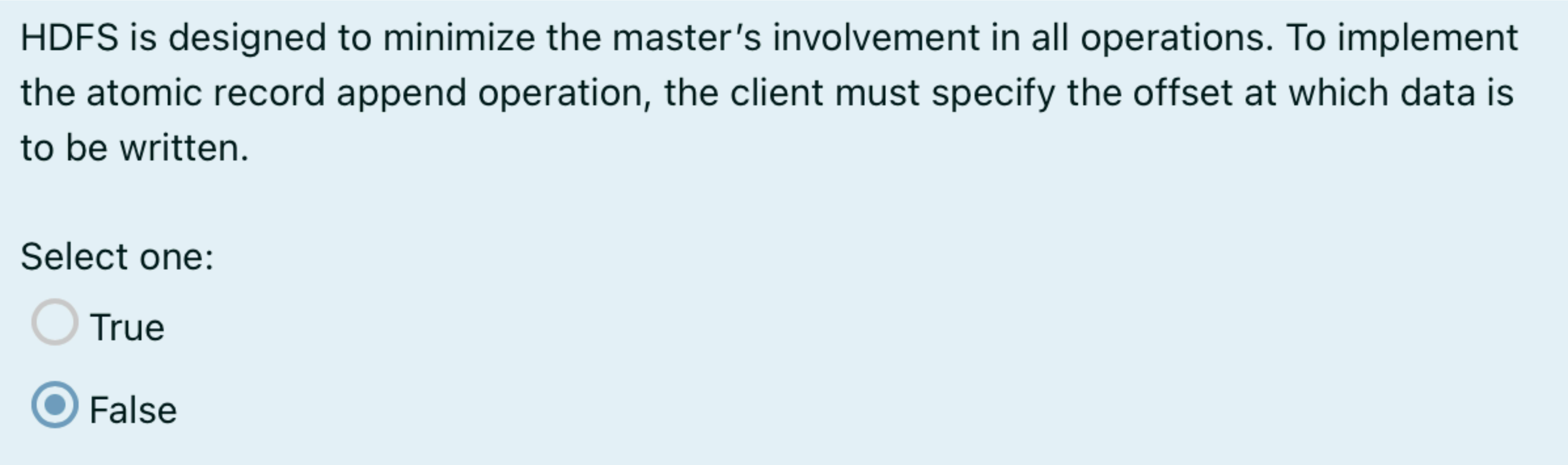


Which of the following can be consider features of HDFS?

- ☑ A. High availability
- ☑ B. Replication
- □ C. Transactional support
- □ D. Real-time
- ☑ E. Fault tolerance

- HDFS is a files system. It does have an edit log which is called Transaction Log but this does not have any relation to Transaction Management in databases. It is log to keep a record of changes that have happened in the file system. However, there are databases built on top of HDFS such as column-oriented non-relational database called HBase, or a data warehouse called Hive that allows you to achieve ACID properties.

Question 6



HDFS is designed to minimize the master's involvement in all operations. To implement the atomic record append operation, the client must specify the offset at which data is to be written.

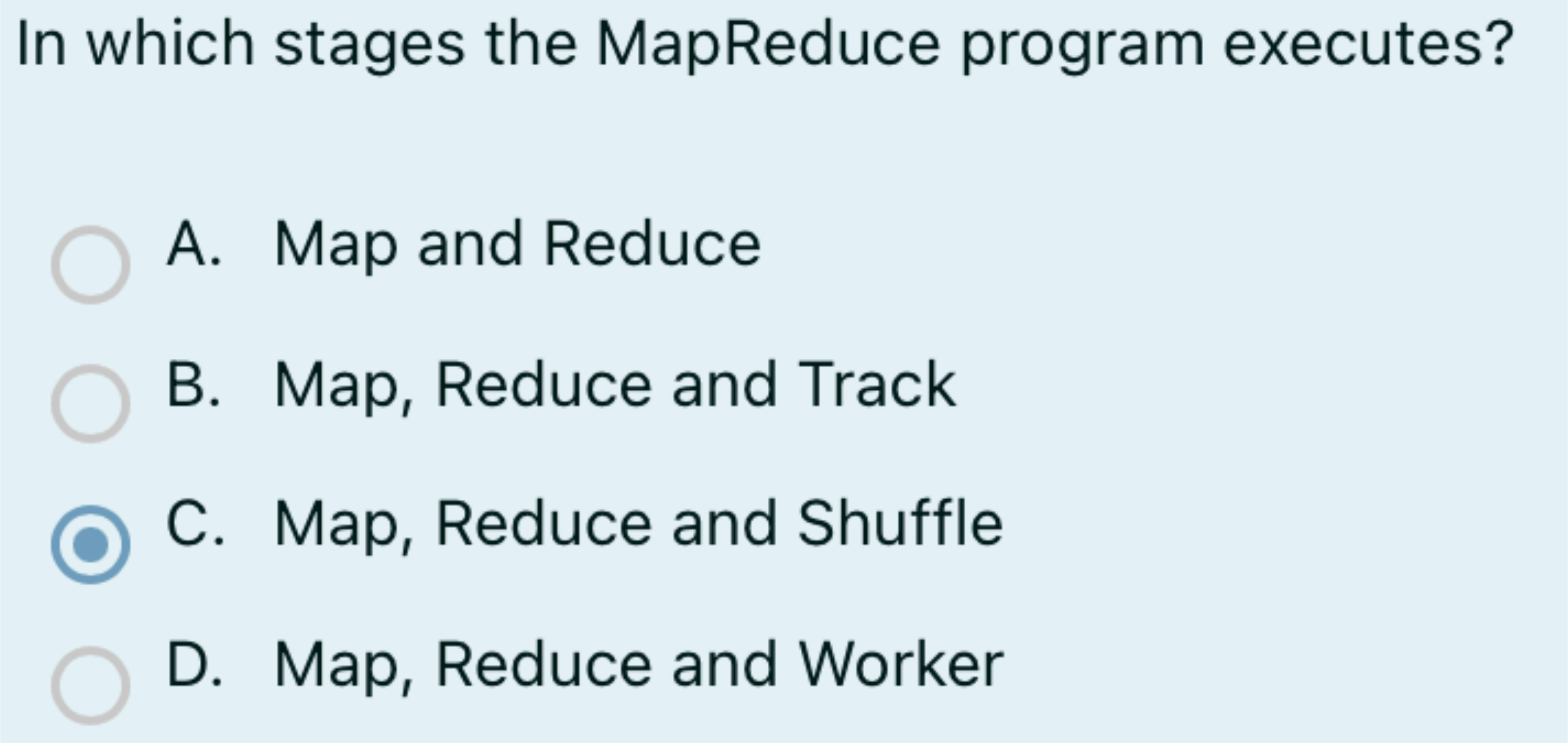
Select one:

○ True

● False

- From the GFS paper, section 3.3 “Atomic Record Appends” - “GFS provides an atomic append operation called record append. In a traditional write, the client specifies the off- set at which data is to be written.”

Question 8



In which stages the MapReduce program executes?

○ A. Map and Reduce

○ B. Map, Reduce and Track

● C. Map, Reduce and Shuffle

○ D. Map, Reduce and Worker

- The question asks in which stages a MapReduce program **executes** (and not in which stages it is programmed)**.** The program executes in three phases: map, shuffle and reduce.