

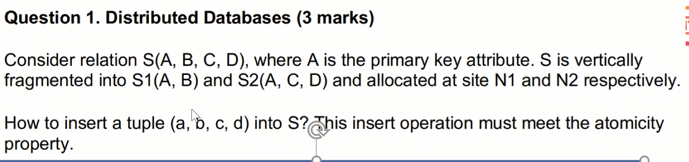
Semi join , horizontally fragmented, 分割S，寻求共同属性并且放于同一个子表中，

R和S的所有分块中都要带有primary Key , R发到site2, S发到site1。计算cost 的结果，比较哪个消耗小。

Approach1: site 2 -- S ---> site1

Approach2: site 1 - all unique value of R.A -> site2

Site 2 - S semijoin(用符号) -> site 1

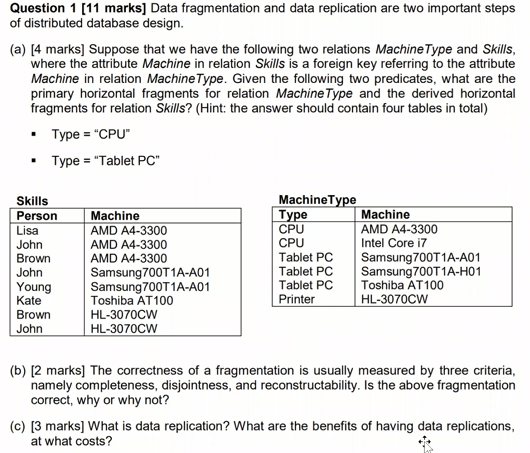


Check if a exists in S (to meet the primary key constraints)

Check with N1 and N2 to see if S1 and S2 can be updated. (if update locks can be granted)

Insert (a,b) to S1 and insert (a,c,d) to S2

Need two PC to guarantee atomicity

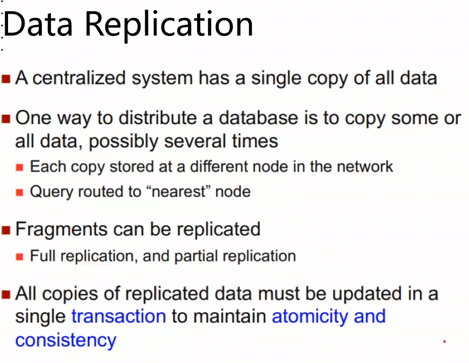


1,type = cpu, 2. type = Tablet PC

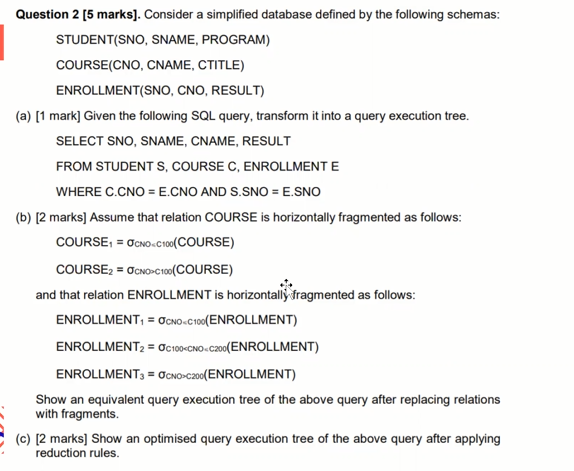
把skill 表中的前三分在一起，456分在一起

1. 出现了疏漏，违反了 reconstruct ability,最后一行没有取到

3，

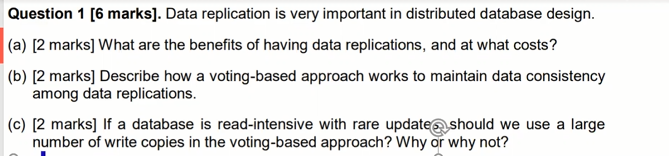


找的属性和划分属性不一样，是不可以直接做reduction的，数据库需要把所有的划分都遍历一遍才可以。



Updating distributed data

Two phase commit protocol



2019 s1

Q1:

Pros:

Accelerated data retrieval when

data is stored locally

Data backup

Cons:

Extra storage cost

Update cost. When the original

data is updated, we need to

update the replicas too.

Suppose we have S replicas. Whenever there is a need to update them, we update the majority copies, say m copies. The version number of each copy is updated at the same time too. When we read, we need to read at least n copies such that n + m > S.

We should adopt the Read-Any Write-All update strategy. Since the database is read-intensive, we need to perform a lot of read operations using the voting strategy. Although we need to write many times in RAWA, the database involves few updates.

Q2:

这个题要注意让写的是哪个query，因为老师说不考画Tree所以，可能会给Tree让写Query

SELECT SNO, SNAME, CNAME, RESULT

FROM STUDENT S, COURSE C, ENROLLMENT E

WHERE C.CNO = E.CNO AND S.SNO = E.SNO

horizontally fragmented as follows: （如何划分范围，这个题可以作为示例）

COURSE1 = σCNO≤C100(COURSE)

COURSE2 = σCNO>C100(COURSE)

或

ENROLLMENT1 = σCNO≤C100(ENROLLMENT)

ENROLLMENT2 = σC100<CNO≤C200(ENROLLMENT)

ENROLLMENT3 = σCNO>C200(ENROLLMENT)

Q3

We must materialize PLT.

Choose one from PL, PT, LT

Let’s say we choose to materialize PL

PLT: 10000 x 0.05

PL: 5000 x 0.25

PT: 10000 x 0.15

LT: 10000 x 0.1

P: 5000 x 0.2

L: 5000 x 0.1

T: 10000 x 0.1

NULL: 5000 x 0.05

Weighted sum

Q4

庞大的数据量

数据质量

不同数据格式的识别。

数据格式转换（转换为业务分析可用的格式）

1. 可以匹配的进行整合，不可以匹配的，进行插入操作
2. SQL query:

Q5:

4个数据质量维度：

例子：

Record linkage 的意义并解释

Edit distance 的计算

Jaccard coefficient 和 Edit distance 的比较联系，什么时候使用

Blocking, Sorted Neighbourhood Approach, Clustering and Canopies, etc.

解释以上的 record linkage 之一

Q6：

K-anonymity is a key concept that was introduced to address the risk of re-

identification of anonymised data through linkage to other datasets.

Generalization: individual values of attributes are replaced with a broader

category. For example, the value “19” of the attribute “age” can be replaced by

“under 20”.

K匿名可能会导致部分匿名处理但是同属性区间下的信息依然可以被查询出来，L匿名的优点就是降低这个问题，方法是L匿名通过重新对区间二次匿名，达到保密的效果

Olap oltp

(DW) (OPERATION)

LOGICAL SOFTWARE

LOW DATA HUGE DATE

FACE TO OBJECT FACE TO ANALYSIS

CAN UPDATE ONLY CAN READ

E-R MODEL STAR/ SKNOWFLAKE MODEL

物化视图在DW中的优缺点

优点：

Pre-calculate expensive joins

提高了查询性能

增删不影响software SQL 语句

缺点：

占用存储空间，存储成本高。

原表更新视图也要更新。

更新消耗大