Kumar Rajesh
201614060112

Data base Techniques

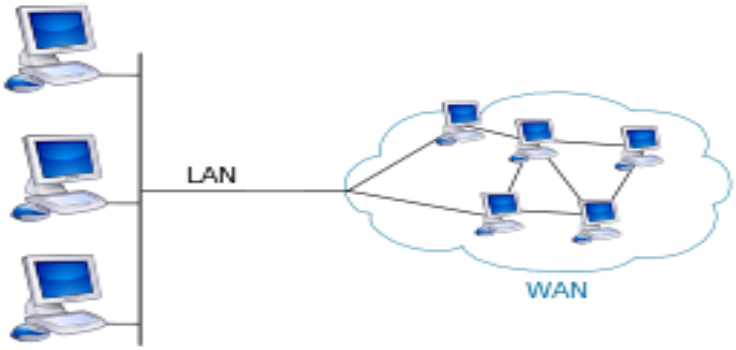
Home Work Two

Kumar Rajesh
11/11/2016

HOMEWORK TWO

Q:1 How might a distributed database designed for a Local-Area Network differ from one designed for a wide-area network?

ANS:



1. Data transfer on LAN is much faster than on WAN. Thus replication and fragmentation will not increase throughput and speed-up on a LAN, as much as in a WAN. But even in a LAN, replication has its uses in increasing reliability and availability.
2. The main difference is range of coverage...A WAN covers the largest area and can cover multiple cities, states, and can span across different continents. An example of a WAN would be the World Wide Web. LANs cover a smaller area...Office networks in the same building or area and home networks are examples LANs

Q:2 To build a highly available distributed system, you must know what kinds of failures can occur.

A. List possible types of failure in a distributed system.

ANS:

i. Computer failure (site failure).

ii. Disk failure.

(iii). Communication failure.

B. Which items in your list from part a are also applicable to a centralized system?

ANS:

(i). Computer failure (site failure).

(ii). Disk failure.

Q:3 Consider a failure that occurs during 2PC for a transaction, For each possible failure that you listed in 2.a, explain how 2PC ensures transaction atomicity despite the failure.

ANS:

A proof that 2PC guarantees atomic commits/aborts in spite of site and link failures, follows.
The main idea is that after all sites reply with a <ready *T*> message, only the co-coordinator of a transaction can make a commit or abort decision. Any subsequent commit or abort by, a site can happen only after it ascertains the co-coordinator's decision, either directly from the coordinator, or indirectly from some other site. Let us enumerate the cases for a site aborting, and then for a site committing.

(A). A site can abort a transaction T (by writing an <abort *T*> log record) only under the following circumstances:-

i. It has not yet written a <ready *T*> log-record. In this case, the coordinator could not have got, and will not get a<ready *T*> or <commit *T*> message from this site. Therefore only an abort decision can be made by the coordinator.

ii. It has written the <ready *T*> log record, but on inquiry it found out that some other site has an <abort *T*> log record. In this case it is 210 Distributed Databases correct for it to abort, because that other site would have ascertained the coordinator's decision (either directly or indirectly) before actually aborting.

iii. It is itself the coordinator. In this case also no site could have committed, or will commit in the future, because commit decisions can be made only by the coordinator.

(B). A site can commit a transaction T (by writing an <commit *T*> log record) only under the following circumstances:-

i. It has written the <ready *T*> log record, and on inquiry it found out that some other site has a <commit *T*> log record. In this case it is correct for it to commit, because that other site would have ascertained the coordinator's decision (either directly or indirectly) before actually committing.
ii. It is itself the coordinator. In this case no other participating site can abort/ would have aborted, because abort decisions are made only by the coordinator.

Q:4 Consider a relation that is fragmental horizontally by plant_number

Employee (name, address, salary, plant_number)

Assume that each fragment has two replicas: one stored at the New York site and one stored locally at the plant site. Describe a good processing strategy for the following queries entered at the San Jose site?

(A) Find all employees at the Boca plant.

(B) Find the average salary of all employees.

(C) Find the highest-paid employee at each of the following sites: Toronto, Edmonton, Vancouver, and Montreal.

(D) Find the lowest-paid employee in the company.

(A)→ Find all employees at the Boca plant.

ANS:

i. Send the query Π*name* (*employee*) to the Boca plant.
ii. Have the Boca location send back the answer.

(B)→ Find the average salary of all employees.

ANS:

i. Compute average at New York.
ii. Send answer to San Jose.

(C)→ Find the highest-paid employee at each of the following sites: Toronto, Edmonton, Vancouver, Montreal.

ANS:

i. Send the query to find the highest salaried employee to Toronto, Edmonton, Vancouver and Montreal.
ii. Compute the queries at those sites.
iii. Return answers to San Jose.

(D)→ Find the lowest-paid employee in the company.

ANS:

i. Send the query to find the lowest salaried employee to New York.
ii. Compute the query at New York.
iii. Send answer to San Jose.

Q:5 Discuss the relative advantages of centralized and distributed database?

ANS:

A distributed database allows a user convenient and transparent access to data which is not stored at the site, while allowing each site control over its own local data. A distributed database can be made more reliable than a centralized system because if one site fails, the database can continue functioning, but if the centralized system fails, the database can no longer continue with its normal operation.

Also, a distributed database allows parallel execution of queries and possibly splitting one query into many parts to increase throughput. A centralized system is easier to design and

implement. A centralized system is cheaper to operate because messages do not have to be sent.

Q:6 Explain how the following differ: fragmentation transparency, replication transparency, and location transparency.

ANS:

A. With fragmentation transparency, the user of the system is unaware of any fragmentation the system has implemented. A user may formulate queries against global relations and the system will perform the necessary transformation to generate correct output.
B. With replication transparency, the user is unaware of any replicated data. The system must prevent inconsistent operations on the data. This requires more complex concurrency control algorithms.
C. Location transparency means the user is unaware of where data are stored. The system must route data requests to the appropriate sites.

Q:7 When is it useful to have replication or fragmentation of data? Explain your answer?

ANS:

Replication is useful when there are many read-only transactions at different sites wanting access to the same data. They can all execute quickly in parallel, accessing local data. But updates become difficult with replication. Fragmentation is useful if transactions on different sites tend to access different parts of the database.

Q:8. Consider the relations

Employee(name, address, salary, plant_number)

Machine(machine_number, type, plant_number)

Assume that the employee relation is fragmental horizontally bu plant_nmuber, and that each fragmental is stored locally at the corresponding plant site. Assume that the machine relation is stored in its entirety at the Armonk site. Describe a good strategy for processing each of the following queries?

(A) Find all employees at the plant that contains machine number 1130?

ANS:

I. Perform Π*plant number* (*σmachine number*=1130 (*machine*)) at Armonk.
II. Send the query Π*name* (*employee*) to all site(s) which are in the result of the previous query.
III. Those sites compute the answers.
IV. Union the answers at the destination site.

(B) Find all employees at plants that contain machine whose type is "milling machine"

ANS:

This strategy is the same as 0.a, except the first step should be to perform
Π*plant number* (*σtype*="milling machine" (*machine*)) at Armonk.

(C) Find all machines at the Almaden plant.

ANS:

I. Perform *σplant number* = *x* (*machine*) at Armonk, where *x* is the plant number for Almaden.
II. Send the answers to the destination site