CS561: Advanced Topics In Database Systems
Spring-2014

Homework 3

Total Points: 60
Release Date: 04/04/2014
Due Date: 04/14/2014 (11:59PM)

Question 1 [20 Points]

Given a relation for customers that consists of the following attributes:

Customer(ID, name, address, city, age)

The relation is distributed over *m* machines.

Assume the following possible partitioning of the data

1- Records are range-based partitioned on the age column, e.g., from age 10 to age 20 will go to one machine, from age 21 to 30 go to second machine, etc.
2- Records are hash-based partitioned on the entire record, i.e., all columns in the record are inputs to the hash function.
3- Records are hash-based partition on the city column, i.e., the input to the hash function is the city column.

Consider the following queries:

1. Select the average age of all customers grouped by the city. That is:

```
Select city, Avg(age)
From Customer
Group By city;
```

2. Eliminate the duplicates from the Customer table based on age and city attributes. That is:

```
Select distinct age, city
From Customer;
```

3. Select customers in city 'Boston' or 'New York' with age above 30. That is:

```
Select  *
From  Customer
Where (city = 'Boston' or city = 'New York')
And age > 30
```

For each query and each possible partitioning, describe a parallel algorithm to execute the query. Give a sequence of steps as given in lectures for the parallel scan or parallel sorting, for example.

Question 2 [20 Points]

Consider joining to relations R(x1, x2, y) and S(y, z1, z2) based on the y column in a distributed database where R is in one location and S is in another location.

1. Extend the semi-join algorithm given in class to join R and S given that we have a selection condition on R.x1. Describe how the semi-join algorithm will work.

2. Assume we have a condition on S.z1 that is very selective (i.e., only few tuples match the condition predicates), describe an efficient join algorithm in this case between R and S.

3. Give a scenario where semi-join is not effective and its cost is more than the cost of shipping one relation to the other.

Question 3 [20 Points]

For two-phase commit technique in distributed databases, answer the following questions:

1. What is the purpose of the two-phase commit? Why is it used?

2. Assume that in Phase 1 of the technique all sites sent to the coordinator 'ready T' and then one site crashed. What will happen for T, will it commit or abort? What will the crashed site do after recovering?

3. Assume that during Phase 1, the coordinator sent out message 'prepare T' to all sites, and then the coordinator crashed. After a while the coordinator is up again. What will be the produced to decide whether to commit or abort T?

Submission Mechanism

Submit electronically using blackboard system.

Late Policy:

We follow the late policy stated on the course website.