answer – 14.1

1605023

given schema,

person (NID, name, street, city, district, DOB, income)

Employee (eID, NID, organization, position, district)

Here,

Person relation has 160 million tuples and Employee relation has 10 million tuples. These tuples are distributed among 64 nodes ($N_1, N_2, ---, N_{64}$). Partitioning is done based on "district" attribute.

The query to be processed:

SQL ⟶
```
select *
from person p, employee e
where p.NID = e.NID
```

We need to explain how the query will be executed using exchange operator with partitioned join.

Explanation

The query will be processed through following steps:

i. $person' = (P_1', P_2', ---, P_{64}') = xchg(person, range, 64, NID)$.

ii. $employee' = (e_1', e_2', ---, e_{64}')$
$= xchg(employee, range, 64, NID)$.

iii. $P_i \bowtie e_i$ at N_i based on "NID"
for $i \in [1, 64]$.

Ans.

answer – 14.2

1605023

given schema,

Person (NID, name, street, city, district, DOB, income)
Employeel eID, NID, organization, position, district, salary).

Here,
person relation has 160 million tuples and employee relation has 10 million tuples. Tuples are distributed among nodes $N_1, N_2, ---, N_{64}$ where partitioning attribute is district.

The query to be processed:

SQL ⟶ select *
from person p, employee e
where p.income > e.salary

We need to explain how this query will be executed using exchange operator with AFR join.

Explanation

Here, person relation is significantly larger than employee relation. so, the query will be processed through asymmetric FR join. Both the relations are initially partitioned based on "district". The person relation will be repartitioned based on "income" by exchange operator at first. Then, exchange operator will broadcast the whole employee relation.
Thus, employee will be replicated at each node $N_1, N_2, --- N_{64}$ containing $P_1', P_2', ---, P_{64}'$ respectively.
Finally, non-equijoin (based on p.income > e.salary) will be applied between P_i' and employee at each node N_i.

Ans.

answer – 15.1

1605023

We need to explain the impact of skewing using the parallel cost model : $T = \max(T_1, T_2, \ldots, T_n)$ for range partition join of $r \bowtie s$.

Explanation

In "range" partitioning join, relations are at first repartitioned based on an attribute by exchange operator and then locally joined at each node.

Assuming, all nodes have similar processing capacity and exchange operator completes range partitioning task with no irregularity.

Then, execution time for each parallel partitioning task will be similar.

If this repartitioning introduces execution skew by making some nodes highly loaded with tuples, then local join at each node will take different time.

And, nodes with highly loaded data will take significant amount of time compared to other nodes.

In parallel cost model,

operational cost = max { execution time for each parallel process }.

As a result, the operational cost for the parallel join will be significantly high.

Ans.

answer – 16·1

1605023

Date :

Some objectives and applications of data analytics for a mobile phone company like TeleTalk, Grameen Phone, etc. are discussed:

objectives

i. customer satisfaction maximization

ii. profit maximization

applications

i. By analyzing customer profile features, billing pattern, calling pattern, and internet usage pattern, peak hours for different services can be learnt and predicted. User specific packages can be generated and offers can be introduced based on these information.

ii. By analyzing past history of sales, future sales can be predicted. The data can be further used to decide what/how much of a product to produce/stock. Also, customers can be targeted for increasing profit.

Ans.

answer – 16.2

1605023

(a) We have to explain the usage of federated database system or mediator system in the content of the management of higher education in Bangladesh by UGC.

Explanation

University Grants Commission (UGC) may generate and maintain different statistics concerning higher education across the country. Decisions will be made and policies will be imposed based on these stats.

Each institution has its own database of students, faculties, etc. which is maintained by various database management systems.

UGC can use federated DB or mediator system to integrate data from all these databases. Federated system supports global query and update whereas mediator system supports common query but not update. But, UGC can use either of the systems as they do not require to update information.

(b) We have to identify some entities and corresponding attributes for global schema for the above scenario.

Entities with Attributes

i. Student (CGPA, district, thana)

ii. Faculty (qualification), etc.

Ans.