Data Science and Database Technology

Homework 4

1. The following relations are given (primary keys are underlined):

   TOURIST(TouristID, Name, BirthDate, CreditCard, Nationality)
   RESORT(ResortCode, CompanyName, Name, City, #Stars)
   AVAILABLE_SERVICES(ResortCode, ServiceName)
   RESERVATION_STAY(TouristID, StayStartDate, ResortCode,
   StayEndDate, Amount, #Adults, #Children, DownPayment)

   Assume the following cardinalities:

   - card(TOURIST)= 10^5 tuples,
     MIN(BirthDate) = 1-1-1940, MAX(BirthDate) = 31-12-1999,
     distinct values of Nationality $\simeq 100$,
   - card(RESORT)= 10^4 tuples,
     distinct values of City $\simeq 100$,
   - card(AVAILABLE_SERVICES)= 10^4 tuples,
     distinct values of ServiceName $\simeq 10$,
   - card(RESERVATION_STAY)= 10^9 tuples,
     MIN(Date) = 01-01-2017, MAX(Date) = 31-12-2017,
     MIN(Amount) = 1, MAX(Amount) = 10,000

   Furthermore, assume the following reduction factor for the group by condition:

   - having COUNT(*)$\geq$2 $\simeq \frac{1}{10}$.

   Consider the following SQL query:

```
select R.RCode, R.Name, SUM(Amount)
from RESORT R, RESERVATION_STAY RS1
where R.RCode=RS1.RCode and R.City='Santo Domingo'
      and RS1.TouristID IN (select RS.TouristID
                      from RESERVATION_STAY RS, AVAILABLE_SERVICES AS
                      where RS.RCode=AS.RCode and RS.StayStartDate ≥ 01/07/2017 and
                     (AS.ServiceName='Wi-Fi' or AS.ServiceName='Tennis court')
                      group by RS.TouristID
                      having COUNT(*)≥2)
group by R.RCode, R.Name
```

   For the SQL query:

   (a) Report the corresponding algebraic expression and specify the cardinality of each node (representing an intermediate result or a leaf). If necessary, assume a data distribution. Also analyze the group by anticipation.

   (b) Select one or more secondary physical structures to increase query performance. Justify your choice and report the corresponding execution plan (join orders, access methods, etc.).