

Data Science And Database Technology

Homework 4

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

```
CLEANING-COMPANY(CId, Name, Address, City, Region)
OFFERED-SERVICES(CId, SId)
SERVICES(SId, ServiceName, Category)
BUILDING(BId, BuildingName, BuildingType, Address, City, Region)
CLEANING-SERVICES(CId, BId, Date, SId, Cost, NumberOfHours)
```

Assume the following cardinalities:

- $\text{card}(\text{CLEANING-COMPANY}) = 10^4$ tuples,
distinct values of Region = 20
- $\text{card}(\text{OFFERED-SERVICES}) = 2 \cdot 10^5$ tuples
- $\text{card}(\text{SERVICES}) = 100$ tuples,
distinct values of Category = 10
- $\text{card}(\text{BUILDING}) = 5 \cdot 10^7$ tuples,
distinct values of City = 1000
distinct values of BuildingType = 10
- $\text{card}(\text{CLEANING-SERVICES}) = 10^9$ tuples,
 $\text{MIN}(\text{Date}) = 1/1/2010$, $\text{MAX}(\text{Date}) = 31/12/2019$

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

- $\text{having COUNT}(\ast) > 1 \simeq \frac{1}{2}$.
- $\text{having SUM}(\text{Cost}) \geq 1000 \simeq \frac{1}{10}$.

Consider the following SQL query:

```
select BId, SUM(Cost) as TotCost, SUM(NumHors) as TotHours
from CLEANING-SERVICES CS, BUILDING B
where CS.Date >= 1/1/2019 and CS.Date <= 31/12/2019
and B.BuildingType <> 'Office'
and B.City = 'Turin'
and CS.BId = B.BId
and CS.SId IN ( select OS.SId
                from CLEANING-COMPANY CC, SERVICES S, OFFERED-SERVICE OS
                where OS.SId = S.SId and OS.CId = CC.CId
                and (Region = 'Piedmont' or Region = 'Liguria')
                and Category = 'IndoorCleaning'
                group by OS.SId
                having COUNT(\ast) > 1)
group by CS.BId
having SUM(Cost) >= 1000
```

Homework tasks

For the SQL query:

1. 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**.
2. **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.).