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:

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

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

- having COUNT(*)>1 $\simeq \frac{1}{2}$.
- having SUM(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(*)>1)
group by CS.BId
having SUM(Cost)>=1000
```

IN -> semijoin
NOT IN -> antisemijoin

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.).