

# Data Science e Tecnologie per le Basi di Dati

## Homework 4

1. Sono date le relazioni seguenti (le chiavi primarie sono sottolineate):

```
UTENTE(UIId, Nome, Email, Continente, Stato, DataNascita)
POST(PIId, PUIId, TimeStamp, Testo, Visibilità, Luogo*)
LIKED(LUIId, PId, TimeStamp, Tipo)
COMMENTO(CUIId, PId, CUIId, TimeStamp, Testo)
```

Vincoli di integrità referenziale:

```
POST(PUIId) REFERENCES UTENTE(UIId)
LIKED(LUIId) REFERENCES UTENTE(UIId)
COMMENTO(CUIId) REFERENCES UTENTE(UIId)
```

Si ipotizzino le seguenti cardinalità:

- $\text{card}(\text{UTENTE}) = 10^9$  tuple,  
Numero di utenti in Europa =  $2 \cdot 10^8$ ,  
 $01/01/1980 \leq \text{DataNascita} \leq 31/12/1999$
- $\text{card}(\text{POST}) = 10^{11}$  tuple,  
 $01/01/2009 \leq \text{TO\_DATE}(\text{TimeStamp}, \text{'DD/MM/YYYY'}) \leq 31/12/2018$
- $\text{card}(\text{LIKED}) = 5 \cdot 10^{12}$  tuples,  
valori distinti di Tipo = 5
- $\text{card}(\text{COMMENTO}) = 5 \cdot 10^{11}$  tuple,  
 $01/01/2009 \leq \text{TO\_DATE}(\text{TimeStamp}, \text{'DD/MM/YYYY'}) \leq 31/12/2018$

Inoltre si ipotizzi il seguente fattore di riduzione per le condizioni di group by:

- $(\text{having COUNT}(\ast) > 100) \simeq \frac{1}{10}$ .
- $(\text{having COUNT}(\ast) > 40) \simeq \frac{1}{2}$ .

Si consideri la seguente query SQL:

```
select UIId, Stato
from USER U, POST P, LIKED L
where U.Continente='Europa' and TO_DATE(P.TimeStamp, 'DD/MM/YYYY') >= 01/01/2018
and L.Tipo = 'Love' and U.UIId = P.PUIId and P.PId = L.PId
and U.UIId IN ( select U.UIId
                from UTENTE U, COMMENTO C
                where U.DataNascita >= 01/01/1993 and U.DataNascita <= 31/12/1994
                and U.Continente = 'Europa'
                and TO_DATE(TimeStamp, 'DD/MM/YYYY') >= 01/01/2018
                and U.UIId = C.CUIId
                group by U.UIId
                having COUNT(\ast) > 40)
group by U.UIId
having COUNT(\ast)>100;
```

Per l'interrogazione SQL

- Si scriva l'espressione algebrica corrispondente, indicando le operazioni svolte, la cardinalità e la selettività di ogni operazione. Dove necessario, si ipotizzi la distribuzione dei dati. Discutere la possibilità di anticipare l'operatore GROUP BY.
- Si scelgano le strutture fisiche accessorie per migliorare le prestazioni dell'interrogazione. Si motivi la scelta e si definisca il piano di esecuzione (ordine e tipo dei join, accesso alle tabelle e/o indici, etc.).