POLITECNICO
DI TORINO

Data Science e Tecnologie per le Basi di Dati

Esame del 21 Febbraio 2023

Valutazione **25,78** su un massimo di 32,00 (**81**%)

Domanda 1

Risposta non data

Punteggio max.: 1,50

1.5 punti (-15% penalità per risposta sbagliata)

La policy di MAX (complete) linkage prevede che la distanza fra due cluster X e Y sia calcolata come:

$$dist(X, Y) = max_{x \in X, y \in Y} dist(x, y)$$

dove $dist(x, y)$ e' una distanza che puo' essere definita fra coppie di punti.

Per un dataset di 5 punti viene calcolata la seguente matrice di distanze.

	a	b	c	d	e
a	0	10	6	5	13
b	10	0	21	12	25
c	6	21	0	4	11
d	5	12	4	0	2
e	13	25	11	2	0

Si applica il clustering gerarchico agglomerativo per estrarre 3 cluster. Viene utilizzata la policy di “MAX linkage” (complete linkage).

Quali sono i 3 cluster ottenuti?

- (a) {a}, {b}, {c, d, e}
- (b) Non e’ possibile rispondere alla domanda con le informazioni a disposizione
- (c) {c, e}, {a, d}, {b}
- (d) {a, d}, {b, e}, {c}
- (e) {d, e}, {a}, {b, c}
- (f) Nessuna delle altre risposte e’ corretta
- (g) {d, e}, {a, b}, {c}
- (h) {b}, {d, e}, {a, c}

Risposta errata.

La risposta corretta è: {b}, {d, e}, {a, c}

Domanda 2

Risposta corretta

Punteggio ottenuto 5,00 su 5,00

5 punti totali (penalità 15% per ogni risposta sbagliata)

Sono date le seguenti tabelle:

Brano(CodB, Titolo, Genere, CodA)
Artista(CodA, Nome, Cognome, Nazionalità, DataNascita)
User(CodU, Nome, Cognome, Nazionalità, DataNascita , Email)
LikeUser(CodU, CodB, Data, Piattaforma)

Sono date le seguenti cardinalità:

- card(BRANO) = 10^8 tuple
  - Valori distinti di Genere = 10
- card(ARTISTA) = $5 \cdot 10^6$ tuple
  - Valori distinti di Nazionalità = 100
  - MIN(DataNascita) = 1/1/1900, MAX(DataNascita) = 31/12/1999
- card(USER) = $2 \cdot 10^7$ tuple
  - MIN(DataNascita) = 1/1/1930, MAX(DataNascita) = 31/12/2004
  - Valori distinti di Nazionalità = 100
- card(LIKEUSER)= 10^{10} tuple
  - Valori distinti di Piattaforma = 5
  - MIN(Data) = 1/1/2003, MAX(Data) = 31/12/2022

Inoltre, sono dati i seguenti fattori di riduzione per le clausole having:

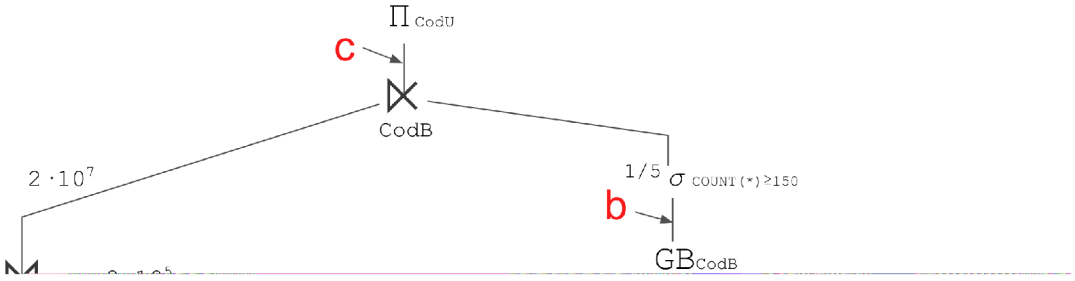
- Having COUNT(*)≥150 = 1/5

Si consideri la seguente query:

```
select LU1.CodU
from  User U, LikeUser LU1
where U.CodU=LU1.CodU
and LU1.Data ≥ 1/1/2022
and U.DataNascita ≥ 1/1/2002
and LU1.CodB IN (SELECT LU2.CodB
  FROM Brano B, Artista A, LikeUser LU2
  WHERE B.CodA=A.CodA and LU2.CodB=B.CodB
  and LU2.Piattaforma='Social'
  and A.DataNascita≥ 1/1/1980
  and B.Genere='Jazz'
  GROUP BY LU2.CodB
  HAVING COUNT(*)≥150
```

Cardinalità

(1.5 punti, penalità -15% per ogni risposta sbagliata)
La figura sottostante rappresenta il query tree per la query precedente.



Seleziona la risposta corretta per la cardinalità d

Seleziona la risposta corretta per la cardinalità di **(b)**:

- 10^8
- 10^7
- $2 \cdot 10^6$ ✓
- 10^4

Punteggio ottenuto 5,00 su 5,00

La risposta corretta è: $2 \cdot 10^6$

Seleziona la risposta corretta per la cardinalità di **(c)**:

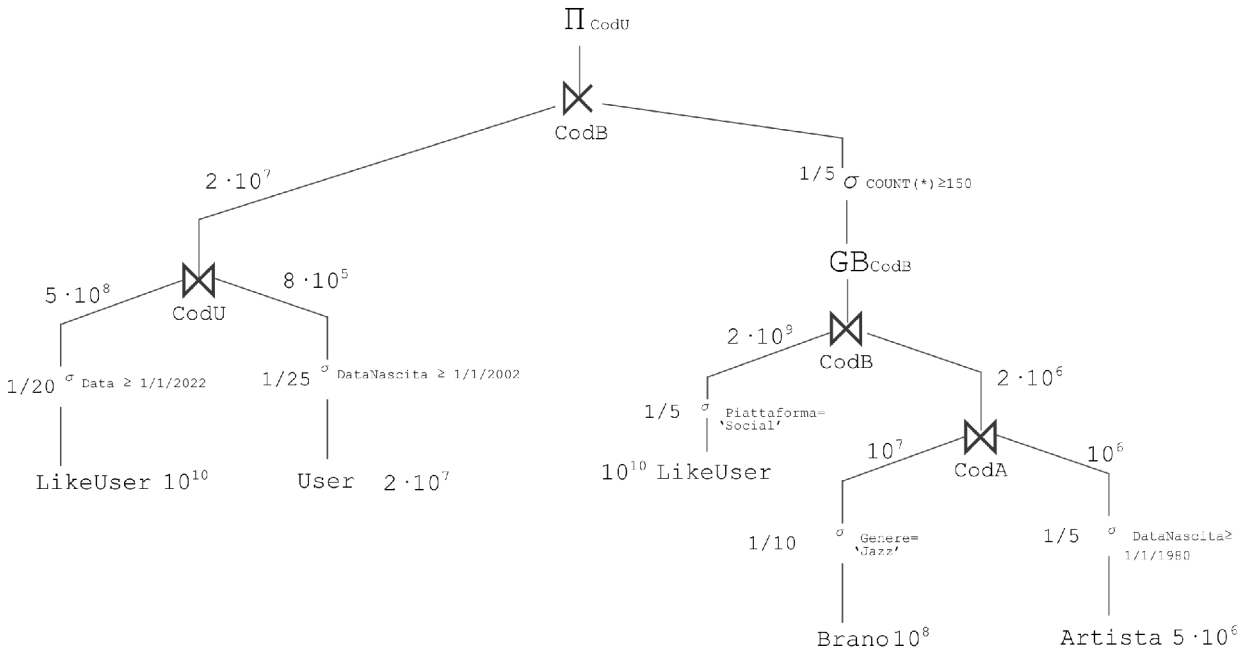
- $8 \cdot 10^6$
- $8 \cdot 10^4$ ✓
- $2 \cdot 10^5$
- $2 \cdot 10^7$

Punteggio ottenuto 5,00 su 5,00

La risposta corretta è: $8 \cdot 10^4$

Indici

(1.5 punti, penalità -15% per ogni risposta sbagliata)
La figura sottostante rappresenta il query tree per la query precedente.



Selezionare, per ogni tabella, le strutture fisiche accessorie per migliorare le prestazioni dell'interrogazione (se possibile) tra le opzioni di seguito.

Tabella LIKEUSER

- (•) CREATE INDEX IndexB ON LIKEUSER(Data) - B+- Tree ✓
- ( ) Nessuna struttura fisica accessoria su questa tabella migliorerebbe le prestazioni dell'interrogazione
- ( ) CREATE INDEX IndexA ON LIKEUSER(Data) - HASH

Punteggio ottenuto 3,00 su 3,00

La risposta corretta è: CREATE INDEX IndexB ON LIKEUSER(Data) - B+- Tree

Tabella LIKEUSER

- ( ) CREATE INDEX IndexD ON LIKEUSER(Piattaforma) - B+- Tree
- ( ) CREATE INDEX IndexC ON LIKEUSER(Piattaforma) - HASH
- (•) Nessuna struttura fisica accessoria su questa tabella migliorerebbe le prestazioni dell'interrogazione ✓

Punteggio ottenuto 3,00 su 3,00

La risposta corretta è: Nessuna struttura fisica accessoria su questa tabella migliorerebbe le prestazioni dell'interrogazione

Tabella USER

- ( ) CREATE INDEX IndexE ON USER(DataNascita) - HASH
- ( ) Nessuna struttura fisica accessoria su questa tabella migliorerebbe le prestazioni dell'interrogazione
- (•) CREATE INDEX IndexF ON USER(DataNascita) - B+- Tree ✓

Punteggio ottenuto 3,00 su 3,00

La risposta corretta è: CREATE INDEX IndexF ON USER(DataNascita) - B+- Tree

Tabella BRANO

- [x] CREATE INDEX IndexG ON BRANO(Genere) - HASH ✓
- [ ] Nessuna struttura fisica accessoria su questa tabella migliorerebbe le prestazioni dell'interrogazione
- [ ] CREATE INDEX IndexH ON BRANO(Genere) - B+- Tree

Punteggio ottenuto 3,00 su 3,00

La risposta corretta è: CREATE INDEX IndexG ON BRANO(Genere) - HASH

Tabella ARTISTA

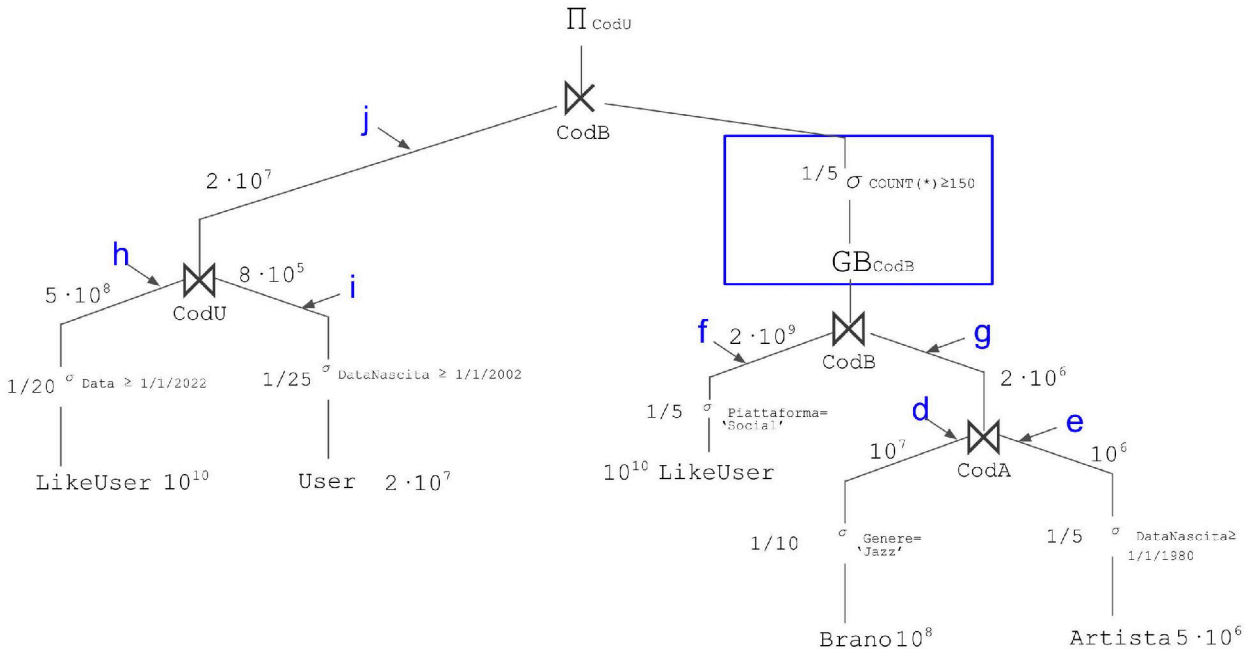
- [ ] CREATE INDEX IndexI ON ARTISTA(Nazionalità) - HASH
- [x] Nessuna struttura fisica accessoria su questa tabella migliorerebbe le prestazioni dell'interrogazione ✓
- [ ] CREATE INDEX IndexJ ON ARTISTA(Nazionalità) - B+- Tree

Punteggio ottenuto 3,00 su 3,00

La risposta corretta è: Nessuna struttura fisica accessoria su questa tabella migliorerebbe le prestazioni dell'interrogazione

Anticipo Group By

(2 punti, penalità -15% per ogni risposta sbagliata)
La figura sottostante rappresenta il query tree per la query precedente.



Analizzare l'anticipo della GROUP BY **GROUP BY LU2.CodB HAVING COUNT(*)≥150** rappresentata nel riquadro. Selezionare la soluzione che **consente la massima efficienza** nell'esecuzione della query (se esiste).

- (•) E' possibile anticiparla nel ramo f ✓
- ( ) E' possibile anticiparla nel ramo d
- ( ) Non è possibile anticipare la Group By GROUP BY LU2.CodB HAVING COUNT(*)≥150
- ( ) E' possibile anticiparla nel ramo e
- ( ) E' possibile anticiparla nel ramo g
- ( ) E' possibile anticiparla nel ramo i
- ( ) E' possibile anticiparla nel ramo j
- ( ) E' possibile anticiparla nel ramo h

Punteggio ottenuto 20,00 su 20,00

La risposta corretta è: E' possibile anticiparla nel ramo f

1) La risposta corretta è : $4 \cdot 10^7$
2) La risposta corretta è : $2 \cdot 10^6$
3) La risposta corretta è : $8 \cdot 10^4$
4) La risposta corretta è : CREATE INDEX IndexB ON LIKEUSER(Data) - B+- Tree
5) La risposta corretta è : Nessuna struttura fisica accessoria su questa tabella migliorerebbe le

prestazioni dell'interrogazione
6) La risposta corretta è : CREATE INDEX IndexF ON USER(DataNascita) - B+- Tree
7) La risposta corretta è : CREATE INDEX IndexG ON BRANO(Genere) - HASH
8) La risposta corretta è : Nessuna struttura fisica accessoria su questa tabella migliorerebbe le prestazioni dell'interrogazione
9) La risposta corretta è : E' possibile anticiparla nel ramo f

Domanda 3

Risposta corretta

Punteggio ottenuto 2,00 su 2,00

2 punti (-15% penalità per risposta sbagliata)

Viene data la seguente base dati transazionale.

	Transactions
0	B C
1	A D E
2	A D E
3	A B C
4	C E
5	B C
6	B D
7	A D E
8	A B C E
9	A E

Si applichi l'algoritmo Apriori per estrarre gli itemset frequenti. Si utilizzi minsup = 2 (un itemset e' frequente se compare in almeno 2 transazioni).

Quali sono gli itemset di lunghezza 3 che vengono generati da Apriori **dopo i passi di join e prune** (con principio Apriori), **prima del conteggio del supporto** nella base dati?

- (a) Nessuna delle altre risposte e' corretta
- (b) ABC, ACE, ADE ✓
- (c) ABC, ABE
- (d) ABC, ADE

- (e) ABD, ABE, ACD, ECD
- (f) ABC, ABD, ABE, ACD, ACE, ADE, ECD
- (g) ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE
- (h) ABC, ABD, ACE, ADE
- (i) Non e' possibile rispondere alla domanda con le informazioni a disposizione

Risposta corretta.

La risposta corretta è: ABC, ACE, ADE

Domanda 4

Risposta corretta

Punteggio ottenuto 1,00 su 1,00

1 punto (penalità 15% per risposta sbagliata)

Il seguente albero mostra un sottoinsieme di una base dati con i diversi livelli di granularità (e.g. tabelle/frammenti/record). Per ogni nodo dell'albero, sono riportati i lock già acquisiti da altre transazioni tra [parentesi quadre].