Basi di dati II

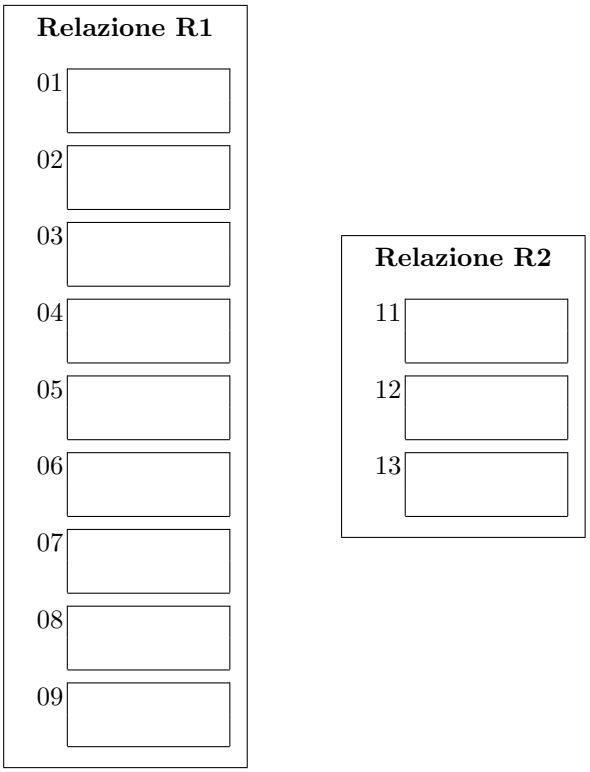
Prova parziale — 9 aprile 2018 — Compito A

Tempo a disposizione: un'ora e quindici minuti.

Cognome _____ **Nome** _____ **Matricola** _____

Basi di dati II — 9 aprile 2018 — Compito A

Domanda 1 (30%)
Considerare le relazioni R1 ed R2 schematizzate sotto. I riquadri interni indicano i blocchi e il numero a fianco a ciascun riquadro indica l'indirizzo del blocco.

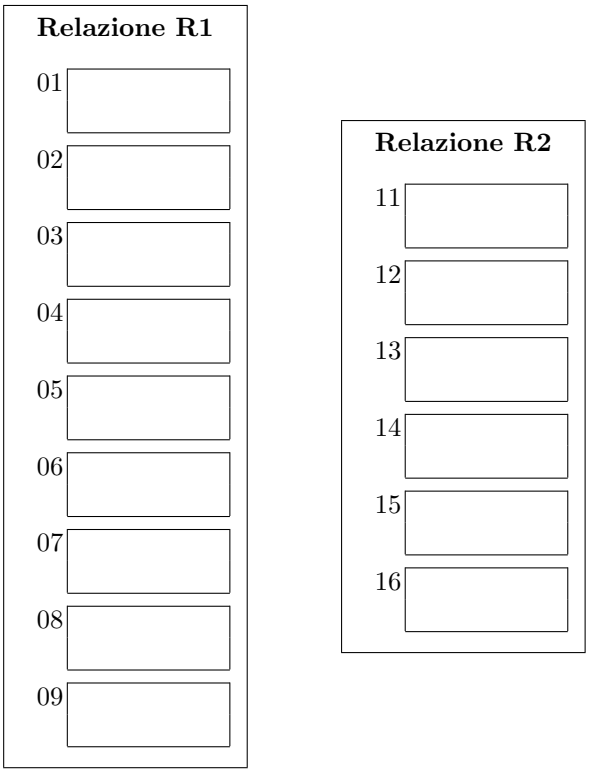


Supponendo di disporre di un buffer di quattro pagine, considerare l'esecuzione di un join con nested loop (senza indici) e rispondere ai quesiti seguenti.

Indicare, nell'ordine, gli indirizzi dei blocchi su cui si eseguono operazioni di pin (o fix) per eseguire l'intero join (riportando più volte gli indirizzi su cui si eseguono eventualmente più pin).

Indicare simbolicamente il costo complessivo dell'operazione, in termini di accessi a blocchi, denotando con L_1 ed L_2, rispettivamente, il numero di blocchi di R1 e R2 e con P il numero di pagine di buffer disponibili.

Considerare ora le relazioni R1 ed R2 schematizzate sotto.



Supponendo di disporre sempre di un buffer di quattro pagine, considerare ancora l'esecuzione di un join con nested loop (senza indici) e rispondere ai quesiti seguenti (uguali a quelli posti nella domanda precedente).

Indicare, nell'ordine, gli indirizzi dei blocchi su cui si eseguono operazioni di pin (o fix) per eseguire l'intero join (riportando più volte gli indirizzi su cui si eseguono eventualmente più pin).

Indicare simbolicamente il costo complessivo dell'operazione, in termini di accessi a blocchi, denotando con L_1 ed L_2, rispettivamente, il numero di blocchi di R1 e R2 e con P il numero di pagine di buffer disponibili

Basi di dati II — 9 aprile 2018 — Compito A

Domanda 2 (25%)
Si considerino un sistema con blocchi di dimensione $B = 4000$ byte e una relazione *R(A,...)* di cardinalità pari circa a $L = 4.000.000$, con ennuple di $e = 100$ byte e campo chiave A di tipo stringa (ad esempio un nome). Valutare la possibile struttura della relazione fra le seguenti alternative:

- struttura disordinata senza indici (nemmeno sulla chiave primaria)
- struttura disordinata con indice sulla chiave primaria A
- struttura hash (dinamica) sulla chiave primaria A

Considerare il seguente carico applicativo:

1. inserimento di una nuova ennupla (con verifica del soddisfacimento del vincolo di chiave), con frequenza $f_1 = 2.000$;
2. ricerca di una ennupla sulla base del valore della chiave A, con frequenza $f_2 = 1.000$;
3. ricerca di una ennupla sulla base del valore parziale (una sottostringa iniziale) della chiave A, con frequenza $f_3 = 10$; supporre che il valore parziale sia molto selettivo e porti alla identificazione, in media, di $s = 5$ ennuple.

Ragionare in termini di numero di accessi a memoria secondaria, assumendo che (1) l'indice abbia profondità $p = 4$, (2) il buffer disponibile permetta di mantenere stabilmente in memoria due livelli dell'indice, (3) l'hash dinamico, per gli accessi puntuali, abbia costo unitario, (4) lettura e scrittura abbiano lo stesso costo.
Rispondere negli spazi sottostanti, in forma sia simbolica sia numerica.

	disordinato senza indice	disordinato con indice	hash dinamico
Costo unit. Op. 1			
Costo unit. Op. 2			
Costo unit. Op. 3			
Costo complessivo			

Basi di dati II — 9 aprile 2018 — Compito A

Domanda 3 (30%)
Si consideri una base di dati con le relazioni (entrambe con indice sulla chiave primaria)

R1(A,B,C), R2(D,E,F)

Eseguendo le interrogazioni seguenti, su Postgres, si rilevano le seguenti scelte per l'operatore di join:

1.	`select *` `from R1 join R2 on C=D`	Hash join

2.	`select *`
	`from R1 join R2 on C=D`
	`where B>=41 AND B<=45`
		Nested loop join
		con accesso diretto alla relazione interna

Motivare ciò, valutando, per ciascuna delle due interrogazioni, il costo di un piano di esecuzione con hash join e uno con nested loop join (e che, per l'operazione 2, includa la selezione), supponendo che

- le relazioni abbiano N_1=2.000.000 ed N_2=4.000.000 ennuple, (con fattore di blocco f_1=20 e f_2=40)
- l'attributo B in R1 abbia circa 200.000 valori diversi (compresi fra 1 e 200.000 e distribuiti uniformemente)
- entrambi gli indici abbiano p=4 livelli (radice e foglie incluse) e fattore di blocco massimo f_i=50
- l'operazione possa contare su un numero di pagine di buffer pari a circa q=500.

Rispondere riempiendo la tabella sottostante, indicando il costo in modo sia simbolico sia numerico.

	Hash join	Nested loop join con accesso diretto ...
1.		
2.		

Indicare come cambiano i costi, per l'operazione 2, se sull'attributo B è definito un indice.

	Hash join	Nested loop join con accesso diretto ...
2bis.		

Domanda 4 (15%)
Si consideri un B-tree con nodi intermedi che contengono due chiavi e tre puntatori e foglie con due chiavi, in cui vengano inserite chiavi (a partire dall'albero vuoto) nel seguente ordine: 51, 57, 11, 32, 14, 27, 28, 31, 34, 35, 36. Mostrare l'albero dopo l'inserimento di cinque chiavi, di otto chiavi e alla fine.

Basi di dati II

Prova parziale — 9 aprile 2018 — Compito A

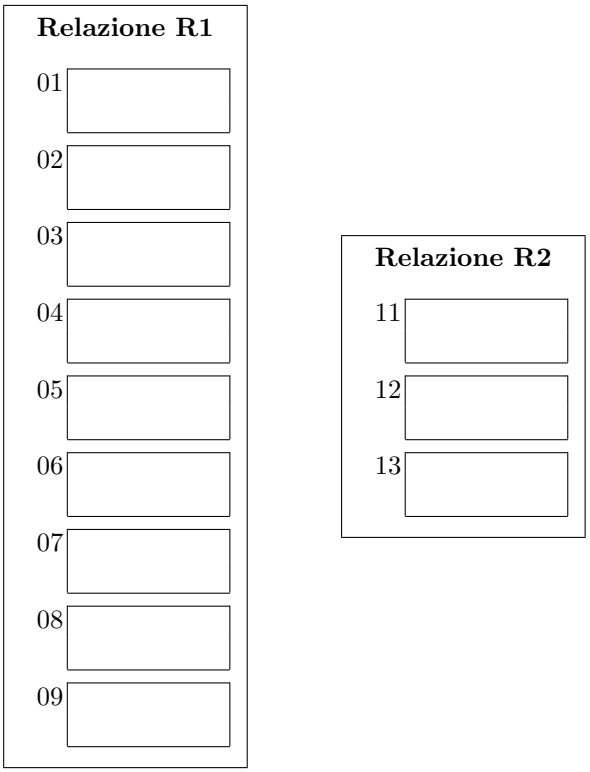
Cenni sulle soluzioni

Tempo a disposizione: un'ora e quindici minuti.

Cognome _____ **Nome** _____ **Matricola** _____

Basi di dati II — 9 aprile 2018 — Compito A

Domanda 1 (30%)
Considerare le relazioni R1 ed R2 schematizzate sotto. I riquadri interni indicano i blocchi e il numero a fianco a ciascun riquadro indica l'indirizzo del blocco.



Supponendo di disporre di un buffer di quattro pagine, considerare l'esecuzione di un join con nested loop (senza indici) e rispondere ai quesiti seguenti.

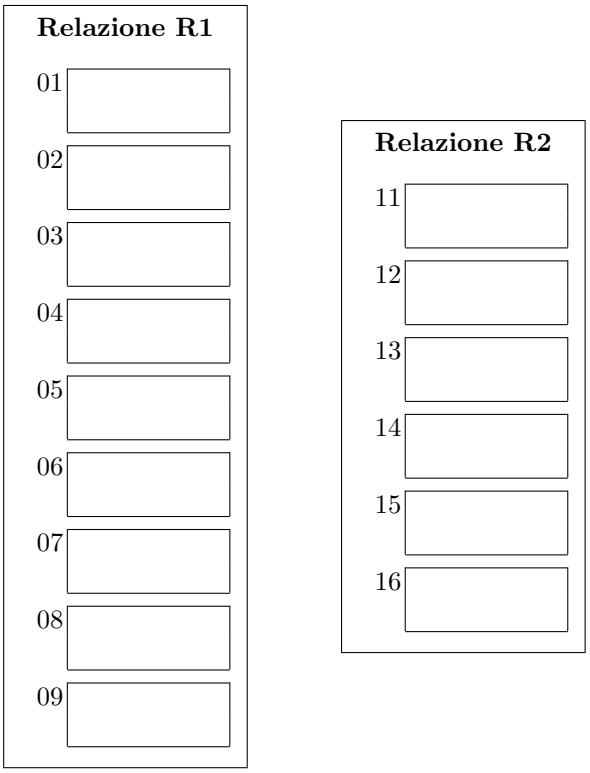
Indicare, nell'ordine, gli indirizzi dei blocchi su cui si eseguono operazioni di pin (o fix) per eseguire l'intero join (riportando più volte gli indirizzi su cui si eseguono eventualmente più pin).

11, 12, 13, 01, 02, 03, ...,09 (per i compiti A e C, per gli altri, invertite con la domanda seguente)

Indicare simbolicamente il costo complessivo dell'operazione, in termini di accessi a blocchi, denotando con L_1 ed L_2, rispettivamente, il numero di blocchi di R1 e R2 e con P il numero di pagine di buffer disponibili.

$L_2 + L_1 = 12$ (idem)

Considerare ora le relazioni R1 ed R2 schematizzate sotto.



Supponendo di disporre sempre di un buffer di quattro pagine, considerare ancora l'esecuzione di un join con nested loop (senza indici) e rispondere ai quesiti seguenti (uguali a quelli posti nella domanda precedente).

Indicare, nell'ordine, gli indirizzi dei blocchi su cui si eseguono operazioni di pin (o fix) per eseguire l'intero join (riportando più volte gli indirizzi su cui si eseguono eventualmente più pin).

11, 12, 13, 01, 02, 03, ...,09, 14, 15, 16, 01, 02, 03, ...,09 (per i compiti A e C)

Indicare simbolicamente il costo complessivo dell'operazione, in termini di accessi a blocchi, denotando con L_1 ed L_2, rispettivamente, il numero di blocchi di R1 e R2 e con P il numero di pagine di buffer disponibili

$L_2 + (L_2/(P-1)) \times L_1 = 24$ (idem)

Basi di dati II — 9 aprile 2018 — Compito A

Domanda 2 (25%)
Si considerino un sistema con blocchi di dimensione $B = 4000$ byte e una relazione *R(A,...)* di cardinalità pari circa a $L = 4.000.000$, con ennuple di $e = 100$ byte e campo chiave A di tipo stringa (ad esempio un nome). Valutare la possibile struttura della relazione fra le seguenti alternative:

- struttura disordinata senza indici (nemmeno sulla chiave primaria)
- struttura disordinata con indice sulla chiave primaria A
- struttura hash (dinamica) sulla chiave primaria A

Considerare il seguente carico applicativo:

1. inserimento di una nuova ennupla (con verifica del soddisfacimento del vincolo di chiave), con frequenza $f_1 = 2.000$;
2. ricerca di una ennupla sulla base del valore della chiave A, con frequenza $f_2 = 1.000$;
3. ricerca di una ennupla sulla base del valore parziale (una sottostringa iniziale) della chiave A, con frequenza $f_3 = 10$; supporre che il valore parziale sia molto selettivo e porti alla identificazione, in media, di $s = 5$ ennuple.

Ragionare in termini di numero di accessi a memoria secondaria, assumendo che (1) l'indice abbia profondità $p = 4$, (2) il buffer disponibile permetta di mantenere stabilmente in memoria due livelli dell'indice, (3) l'hash dinamico, per gli accessi puntuali, abbia costo unitario, (4) lettura e scrittura abbiano lo stesso costo.
Rispondere negli spazi sottostanti, in forma sia simbolica sia numerica.

	disordinato senza indice	disordinato con indice	hash dinamico
Costo unit. Op. 1	L/(B/e) = 100.000	$(p-2)+1+2=5$ Potrei dover scrivere foglia e blocco dati dopo averli letti	1+1
Costo unit. Op. 2	L/(B/e) = 100.000	$(p-2)+1=3$	1
Costo unit. Op. 3	L/(B/e) = 100.000	$(p-2)+5=7$	L/(B/e) = 100.000
Costo complessivo	ca 3×10^8	ca 13.000	ca 1.000.000

Basi di dati II — 9 aprile 2018 — Compito A

Domanda 3 (30%)
Si consideri una base di dati con le relazioni (entrambe con indice sulla chiave primaria)

R1(A,B,C), R2(D,E,F)

Eseguendo le interrogazioni seguenti, su Postgres, si rilevano le seguenti scelte per l'operatore di join:

1.	`select *` `from R1 join R2 on C=D`	Hash join

2.	`select *`
	`from R1 join R2 on C=D`
	`where B>=41 AND B<=45`
		Nested loop join
		con accesso diretto alla relazione interna

Motivare ciò, valutando, per ciascuna delle due interrogazioni, il costo di un piano di esecuzione con hash join e uno con nested loop join (e che, per l'operazione 2, includa la selezione), supponendo che

- le relazioni abbiano N_1=2.000.000 ed N_2=4.000.000 ennuple, (con fattore di blocco f_1=20 e f_2=40)
- l'attributo B in R1 abbia circa 200.000 valori diversi (compresi fra 1 e 200.000 e distribuiti uniformemente)
- entrambi gli indici abbiano p=4 livelli (radice e foglie incluse) e fattore di blocco massimo f_i=50
- l'operazione possa contare su un numero di pagine di buffer pari a circa q=500.

Rispondere riempiendo la tabella sottostante, indicando il costo in modo sia simbolico sia numerico.

	Hash join	Nested loop join con accesso diretto
1.	Con i buffer disponibili (che sono in numero maggiore della radice quadrata del numero di blocchi del più piccolo dei due file), l'hash-join si può eseguire in due passate, con un costo: $3 \times (\frac{N_1}{f_1} + \frac{N_2}{f_2}) = 600.000$	Per il compito A: $\frac{N_1}{f_1} + N_1 \times (p - 2 + 1) =$ ca. 6.000.000
2.	Si può supporre che la selezione su B produca (vista l'ipotesi sul numero di valori diversi) $b = 50$ ennuple. La selezione richiede una scansione della prima relazione che produce le b ennuple su cui effettuare il join. L'hash join può venire poi eseguito con una singola scansione della seconda relazione (avendo in buffer il risultato della selezione): $(\frac{N_1}{f_1} + \frac{N_2}{f_2}) = 200.000$	Scansione di R1 con selezione e poi accesso diretto a R2 con i valori di delle $b = 50$ ennuple selezionate. Il costo è dominato dalla scansione: $\frac{N_1}{f_1} + b \times (p - 1 + 1) =$ ca. 100.000

Indicare come cambiano i costi, per l'operazione 2, se sull'attributo B è definito un indice.

	Hash join	Nested loop join con accesso diretto ...
2bis.	Accesso diretto per la selezione e poi scansione di R2 (il cui costo domina). Costo: $(p + b) + \frac{N_2}{f_2} =$ ca. 100.000	Accesso diretto a R_1 per la selezione e poi accesso diretto a R2. Costo: $((p + b) + b \times (p - 1 + 1)) =$ ca.200

Basi di dati II — 9 aprile 2018 — Compito A

Domanda 4 (15%)
Si consideri un B-tree con nodi intermedi che contengono due chiavi e tre puntatori e foglie con due chiavi, in cui vengano inserite chiavi (a partire dall'albero vuoto) nel seguente ordine: 51, 57, 11, 32, 14, 27, 28, 31, 34, 35, 36. Mostrare l'albero dopo l'inserimento di cinque chiavi, di otto chiavi e alla fine.

Vengono mostrate le soluzioni per il compito A. Le altre sono analoghe (isomorfe).