Basi di dati II — 22 febbraio 2018

Tempo a disposizione: due ore.

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

Domanda 1 (25%) Considerare un sistema che utilizzi blocchi di lunghezza $D = 4$ KB (approssimabili a 4000 byte) e una tabella R con una struttura fisica heap con record a lunghezza fissa che occupano $L = 20$ byte ciascuno, in cui vengono inserite $M = 100.000$ ennuple, con valori della chiave tutti diversi fra loro e da quelli già nella relazione (quindi il sistema verifica il soddisfacimento del vincolo di chiave e ammette tutte le operazioni).
Rispondere alle domande seguenti, indicando formule e valori numerici:

Indicare il numero di scritture di blocchi in memoria secondaria necessarie per realizzare i 100.000 inserimenti, supponendo che i record di log abbiano una lunghezza pari a circa il triplo di quella dei record del file, con riferimento ad un programma che utilizzi una transazione separata per ciascun inserimento (supporre per semplicità che non ci siano altre transazioni attive)

- numero di scritture di pagine di log:

- numero di scritture di pagine della relazione, nei tre casi seguenti:
  - strategia undo-redo senza vincoli particolari:
  - strategia undo-only (no-redo):
  - strategia redo-only (no-undo):

Come nel caso precedente, ma con riferimento ad un programma che, per realizzare i 100.000 inserimenti, utilizzi complessivamente $k = 1000$ transazioni, ognuna con 100 inserimenti (supporre di nuovo che non ci siano altre transazioni attive)

- numero di scritture di pagine di log:

- numero di scritture di pagine della relazione, nei tre casi seguenti:
  - strategia undo-redo senza vincoli particolari:
  - strategia undo-only (no-redo):
  - strategia redo-only (no-undo):

Basi di dati II — 22 febbraio 2018 — Compito A

Domanda 2 (25%) In Postgres (e, con sintassi diverse, negli altri sistemi) è possibile ordinare fisicamente una relazione con il comando `CLUSTER`. L'ordinamento viene specificato sulla base di un indice già definito sulla stessa relazione. Successive operazioni di inserimento e modifica non rispettano però l'ordinamento, che quindi degenera (per esempio, si può immaginare che gli inserimenti vengano eseguiti aggiungendo i dati in coda al file). L'ordinamento può poi essere ripristinato con un nuovo comando `CLUSTER`.
Sia data una relazione $R(\underline{A}, B, C)$ contenente circa $N = 10.000.000$ ennuple di $r = 100$ Byte ciascuna, con $v = 100.000$ valori diversi per l'attributo C, uniformemente distribuiti (quindi si può supporre che per ogni valore di C ci siano $N/v = 100$ ennuple con tale valore). Supporre che i blocchi abbiano dimensione $B = 4$KB approssimabile come 4.000.
Supporre che vengano eseguite, in sequenza, le operazioni sotto elencate e indicare, nei riquadri, i costi delle `SELECT`:

- `CREATE INDEX RCIX ON R (C)` (creazione di un indice su C; supporre che abbia profondità $p = 4$)
- `CLUSTER R USING RCIX` (ordinamento di R sulla base dell'indice e quindi sull'attributo C)
- `SELECT * FROM R WHERE C = 100`

- inserimento di $N/10 = 1.000.000$ ennuple, in ordine casuale, con valori di C pure uniformemente distribuiti (quindi si può supporre che vengano inserite 10 ennuple per ogni valore)
- `SELECT * FROM R WHERE C = 100`

Rispondere poi alla seguente domanda, sempre con riferimento allo scenario sopra discusso

- Indicare (con una breve motivazione) quale potrebbe essere il costo per l'ordinamento iniziale, supponendo di avere a disposizione buffer per circa 4GB.

Basi di dati II — 22 febbraio 2018 — Compito A

Domanda 3 (15%) Si supponga di dover eseguire una interrogazione che calcola statistiche su una base di dati (ad esempio: trovare per ogni corso la media dei voti assegnati). Indicare (con un breve commento, non più di tre righe) quale livello di isolamento (`READ UNCOMMITTED`, `READ COMMITTED`, `REPEATABLE READ` o `SERIALIZABLE`) si potrebbe scegliere in ciascuno dei seguenti casi

1. per tutti i corsi sono già presenti molti esami e l'interrogazione è eseguita mentre vengono inseriti alcuni esami, comunque pochi per corso (rispetto a quelli già presenti); sono accettabili risultati "approssimati"
2. per tutti i corsi sono già presenti molti esami e l'interrogazione è eseguita mentre vengono inseriti alcuni esami, comunque pochi per corso (rispetto a quelli già presenti); *non* sono accettabili risultati "approssimati"
3. per tutti i corsi sono già presenti molti esami e l'interrogazione è eseguita mentre vengono corretti (cioè modificati) i voti di alcuni esami, comunque pochi per corso (rispetto a quelli già presenti) e senza inserirne di nuovi; *non* sono accettabili risultati "approssimati"
4. per tutti i corsi sono già presenti molti esami e l'interrogazione è eseguita mentre vengono corretti (cioè modificati) i voti di alcuni esami, comunque pochi per corso (rispetto a quelli già presenti) e senza inserirne di nuovi; sono accettabili risultati "approssimati"

1.	2.	3.	4.

Basi di dati II — 22 febbraio 2018 — Compito A

Domanda 4 (35%) Si consideri la seguente porzione dello schema dell'archivio delle carriere degli studenti di una anagrafe ministeriale (supporre che ogni studente abbia un ID diverso per ogni "carriera", cioè per ogni corso di studio cui si iscrive):

- Studenti(CodiceFiscale, Cognome, Nome, DataNascita, TipoMaturità)
- Carriera(IDCarriera, CodiceFiscale, CorsoDiStudio, AnnodiImmatricolazione)
- Iscrizioni(IDCarriera, AnnoAccademico, AnnoDiCorso)
- CorsiDiStudio(CodiceCdS, Titolo, Livello, Classe, Università)
- Lauree(IDCarriera, Data, Voto)

Progettare uno schema dimensionale che permetta di rispondere, fra le altre, alle seguenti interrogazioni:

- calcolare il numero di studenti (con la relativa media dei voti) che si sono laureati in un certo corso di studio (inteso come corso di studio presso una università) in un certo anno accademico (si supponga che, per la data di laurea, l'unico dettaglio rilevante sia l'anno accademico e che esista un modo univoco per associare un anno accademico ad una data di laurea)
- calcolare il numero di laureati per una classe di corsi studio, distinto per tipo di maturità e per numero di anni impiegati per conseguire il titolo (ad esempio, 2, 3, 4, 5, 6, più di 6)
- calcolare il numero di laureati per una classe di corsi studio, distinto per "età alla laurea" (ad esempio, 21, 22, 23, ...26, più di 26)

Assumere che, per ragioni di privatezza e di compattezza, sia opportuno limitare la cardinalità della tabella dei fatti, a patto di permettere la risposta alle precedenti interrogazioni.

Mostrare lo schema dimensionale, specificando la grana scelta.

Basi di dati II — 22 febbraio 2018 — Compito A

Descrivere, informalmente, ma in modo strutturato e comprensibile, il processo di ETL che porta alla tabella dei fatti mostrata in risposta alla domanda precedente