به نام خداوند بخشنده مهربان

طراحی پایگاه داده ها

دانشکده مهندسی کامپیوتر

مریم رمضانی
بهار ۱۴۰۳

تمرین چهارم

تاریخ انتشار: ۱۸ خرداد ۱۴۰۳

۱. پرسش‌های خود درمورد این تمرین را در سامانه کوئرا مطرح کنید.

۲. **سیاست ارسال با تاخیر:** شما در مجموع در طول نیم‌سال می‌توانید از ۹ روز تاخیر استفاده کنید. **همچنین هر تمرین را می‌توانید تا حداکثر ۴ روز با تاخیر تحویل دهید.** تاخیرها با مقیاس ساعت محاسبه شده و به بالا گرد می‌شوند.

۳. **سیاست مشارکت دانشجویان در حل کردن تمارین:** دانشجویان می‌توانند در حل تمارین برای رفع ابهام و یا به‌دست آوردن ایده‌ی کلی با یک‌دیگر مشورت و همفکری کنند. این کار مورد تایید و تشویق تیم ارائه‌ی درس می‌باشد؛ چرا که هم‌فکری و کار گروهی می‌تواند موجب تقویت یادگیری شود. اما به‌دست آوردن جزئیات راه‌حل و نگارش پاسخ باید تماما توسط خود دانشجو انجام شود. **حتما در انتهای پاسخ‌های ارسالی خود نام افرادی که با آن‌ها همفکری کردید را ذکر کنید.**

۴. برای سوال ششم، کد های SQL هر بخش را با نام گذاری مناسب، به طور مثال برای بخش اول 6-1.SQL، را در فایل zip پاسخ خود قرار دهید.

سوالات تئوری (۱۱۰ نمره)

تاریخ تحویل: ۲۹ خرداد ۱۴۰۳

پرسش ۱ (۱۰ نمره) فرض کنید جدول $R(A, B, C, D, E)$ مفروض است. آن را به سه رابطه جزئی تر $S_1(A, B, C)$ ، $S_2(B, C, D)$ و $S_3(A, C, E)$ تجزیه کرده ایم. در هر بخش فرض کنید مجموعه FD های معرفی شده برقرار است و نشان دهید آیا تجزیه با این فرض از وابستگی های تابعی، lossless است یا خیر.

(آ) $B \to E, CE \to A$

(ب) $A \to D \to E, B \to D$

(ج) $AC \to E, BC \to D$

(د) $A \to D, CD \to E, E \to D$

پرسش ۲ (۲۰ نمره) در یک کار آماری، اطلاعات تیم های شرکت کننده در لیگ قهرمانان اروپا بین سال های ۲۰۰۲ تا ۲۰۱۹ جمع آوری شده است. همه اطلاعات موجود در یک جدول به صورت زیر ذخیره شده است :

UCL(year, team, group, winner, final-host-stadium, qualified, knockout-stage, best-scorer, most-assists, attendance-prize)

- کلید اصلی رابطه را $(year, team)$ در نظر بگیرید.
- صفت بولین qualified نشان میدهد آیا تیم از مرحله گروهی صعود کرده است یا خیر.
- صفت knockout-stage آخرین مرحله حذفی است که تیم در صورت صعود از گروه به آن رسیده است. اگر یک تیم از گروه صعود نکرده باشد، این صفت مقدار ثابت groupstage را اخذ کرده است.
- صفت attendance-prize مقدار جایزه ای است که در نهایت به تیم اهدا میشود. تیم هایی که از گروه صعود نکرده اند مقدار ثابت X دریافت میکنند و به ازای هر مرحله بالاتر رفتن در مراحل حذفی، مقدار Y به این جایزه اضافه میشود.
- فینال هر دوره، تنها در یک استادیوم برگزار میشود.
- صفات بهترین گلزن و بیشترین پاس گل، مربوط به هر تیم هستند.
- فرض کنید هیچ صفتی مقدار null ندارد.

با توجه به این پایگاه داده، به سوالات زیر پاسخ دهید.

(آ) توضیح دهید پایگاه داده معرفی شده میتواند دارای چه انواعی از آنومالی در هنگام درج یا حذف باشد. میتوانید این آنومالی ها را با یک مثال نیز توضیح دهید.

(ب) یه کمک شناسایی وابستگی های تابعی ، پایگاه داده معرفی شده را تا NF۲ نرمال سازی کنید.

(ج) به کمک شناسایی وابستگی های ترایا ، پایگاه داده مرحله قبل را تا NF۳ نرمال سازی کنید.

(د) آیا پایگاه داده به دست آمده در مرحله قبل BCNF است ؟ اگر جواب شما خیر است، نرمال سازی تا این مرحله را انجام دهید.

پرسش ۳ (۱۵ نمره)

درستی یا نادرستی هر یک از گزاره‌های زیر را با ذکر دلیل مشخص کنید.

(آ) هر رابطه ای که فقط یک خصیصه با مقادیر غیر تکراری دارد حتماً در BCNF واقع است.

(ب) نرمالسازی عموما به افزایش سرعت کوئری ها منجر میشود.

(ج) نرمالسازی تا سطح NF۳ عموما حجم پایگاه داده را کاهش میدهد.

پرسش ۴ (۱۵ نمره) فرض کنید یک جدول پایگاه داده داریم که از ستون‌های A, B, C, D, E تشکیل شده است. فرض کنید روابط وابستگی تابعی در این جدول به صورت زیر می‌باشند

FD = AE -> BC, AC -> D, CD ->BE, D -> E

با این فرضیات به پرسش‌های زیر پاسخ دهید.

(آ) سه کلید کاندید ارائه دهید و توضیح دهید که چرا این کلیدها کلید کاندید محسوب می‌شوند. (اگر بیش از سه کلید کاندید وجود دارد، ۳ کلید کاندید به دلخواه خود بنویسید.)

(ب) آیا این جدول با این وابستگی‌های تابعی، در فرم BCNF قرار دارد؟ توضیح دهید.

اکنون جدول جدیدی برای رابطه ای به نام R با ستون‌های A, B, C, D در نظر بگیرید. وابستگی‌های تابعی این جدول شامل موارد زیر می‌شوند با توجه به این رابطه به بخش‌های بعد پاسخ دهید.

FD = A -> B, B -> C, C -> D

(آ) اگر این رابطه را به سه رابطه زیر تجزیه کنیم، آیا این تجزیه lossless است یا خیر؟ توضیح دهید.

R1(A, B)
R2(C, D)
R3(A, C)

(ب) این رابطه را به روابطی تجزیه کنید تا به فرم BCNF درآید. در صورتی که در بخش آ پاسخ‌تان بله بوده است، تجزیه ارائه شده در این قسمت باید متفاوت از تجزیه ارائه شده در بخش آ باشد

پرسش ۵ (۱۰ نمره) با توجه به مفاهیم مرتبط با index به سوالات زیر پاسخ دهید.

(آ) تفاوت شاخص های $B-Tree$ و B^+-Tree را تحقیق کرده و شرح دهید و برای هر یک مثالی کاربردی بزنید.

(ب) جدول Employee (id: integer, name: String, salary: integer) را در نظر بگیرید. حال برای کوئری های زیر بگویید که کدام نوع شاخص از بین hash index و $B^+-Tree\ index$ مناسب است.

i.
```
SELECT name,
FROM Employee,
WHERE salary >= 5000
```

ii.
```
SELECT name,
FROM Employee,
WHERE id = 50
```

iii.
```
SELECT name,
FROM Employee,
WHERE id = 50 and salary >= 5000
```

پرسش ۶ (۴۰ نمره) برای این سوال، فایل کوئری، اسکرین شات نتایج و نوشتن تحلیل فرد لازم است. جدولی به اسم transaction با ستون های زیر را برای تراکنش های مالی بسازید. سپس با استفاده از کد پیوست تمرین، ۱.۵ میلیون رکورد را در آن وارد نمایید.

ستون	نوع داده	توضیح
transaction_id	serial PRIMARY KEY	شناسه منحصر به فرد تراکنش
transaction_date	date	تاریخ تراکنش
account_id	int	شناسه حساب مربوط به تراکنش
amount	numeric	مبلغ تراکنش
category	varchar(50)	دسته‌بندی تراکنش
status	varchar(20)	وضعیت تراکنش

(آ) در زیر چهار حالت partitioning به همراه ستون پیشنهادی برای آن روش آورده شده است. کوئری‌های لازم برای ساخت چهار جدول زیر را نوشته و اجرا نمایید تا داده‌ی جدول اصلی در این چهار جدول زیر قرار گیرد.

ستون مورد استفاده	نام جدول مقصد	نام روش
تاریخ تراکنش با تفکیک هر یک سال	transaction_date	range
شناسه حساب با تفکیک از طریق باقی مانده بر ۴	transactions_hash	hash
دسته بندی تراکنش ها	transactions_list	list
تاریخ تراکنش و وضعیت تراکنش	transactions_composite	composite

حال برای پرسش‌های زیر کوئری‌های مناسب برای اجرا بر روی این پنج جدول را نوشته و با استفاده از دستور EXPLAIN ANALYZE سرعت اجرا و عملکرد پیدا کردن آن توسط DBMS را در یک جدول مقایسه‌ای برای هر ۵ جدول فوق یادداشت کرده و تحلیل خود را شرح دهید؟ در هر سوال زیر کدام جدول سرعت پاسخگویی بهتری دارد؟ چرا؟

(ب) تعداد تراکنش‌هایی که در سال ۲۰۲۳ اتفاق افتاده است را برای جدول اصلی و هر جدول پارتیشن‌بندی شده بیابید.

(ج) جمع کل مبلغ تراکنش‌هایی که برای یک شناسه حساب خاص (مثلاً ۱۲۳) در جدول اصلی و هر جدول پارتیشن‌بندی شده اتفاق افتاده است را پیدا کنید.

(د) تعداد تراکنش‌ها برای هر دسته‌بندی را در جدول اصلی و هر جدول پارتیشن‌بندی شده بیابید.

(ه) تعداد تراکنش‌های تکمیل شده را در سال ۲۰۲۳ برای جدول اصلی و هر جدول پارتیشن‌بندی شده پیدا کنید.

(و) میانگین مبلغ تراکنش‌ها برای دسته‌بندی Groceries را در جدول اصلی و هر جدول پارتیشن‌بندی شده پیدا کنید.

(ز) میانگین مبلغ تمامی تراکنش‌ها را برای جدول اصلی و هر جدول پارتیشن‌بندی شده پیدا کنید.

(ح) یک index در جدول اصلی transaction بر روی ستون «تاریخ تراکنش» ایجاد نمایید. سپس کوئری‌هایی بنویسید که در جدول‌های transaction و transaction_range این پرسش را پاسخ دهد که «تعداد تراکنش‌هایی که در سال ۲۰۲۳ با مبلغ بیشتر از ۵۰۰۰ اتفاق افتاده است را برگرداند». سرعت رسیدن به جواب در این دو جدول را باهم مقایسه نمایید. چه نتیجه‌ای می‌گیرید؟