

به نام خدا



دانشگاه تهران

دانشکده فنی

دانشکده مهندسی برق و کامپیوتر



درس داده کاوی

تمرین تشریحی ۲

فروردین ماه ۱۴۰۱

* فهرست

سؤال ۱.....	۳
سؤال ۲.....	۴
سؤال ۳.....	۵
سؤال ۴.....	۶
ملاحظات (حتما مطالعه شود).....	۷

سؤال ۱

یک انبارداده مربوط به واردات محصولات در گمرک را در نظر بگیرید. هر یک از رکوردهای اطلاعاتی ما شامل داده‌های زیر است: تاریخ، جنس وارد شده، وزن، ارزش، محل ورود و کشور صادرکننده.

الف) با توجه به پرس‌وجوهای زیر یک مدل ذخیره‌سازی (ستاره، برف‌دانه یا صور فلکی) برای داده‌های مربوط به واردات ارائه دهید. (تنها در صورت نیاز می‌توانید برای هر پرس‌وجو مدل جدا ارائه نمایید)

i. میانگین وزن واردات گندم در سال ۱۴۰۰ از یک گمرک خاص در هر ماه

ii. ارزش کل واردات از چین از شهریور ۱۴۰۰ تا کنون

iii. کالایی که ارزش واردات آن در اسفند ۱۴۰۰ بیشتر از سایر کالاها بوده است.

iv. ماهی از سال ۱۴۰۰ که بیشترین میزان واردات گندم از نظر وزن را داشته است.

ب) فرض کنید date cube کاملی برای داده‌های مربوط به واردات ایجاد شده است. برای هر یک از پرس‌وجوهای بالا عملیات OLAP مورد نیاز برای پاسخگویی به آن را بیان کنید.

سؤال ۲

فرض کنید یک cuboid پایه ۱۰ بعدی دارای چهار سلول پایه زیر باشد:

$$(a_1, a_2, b_3, a_4, \dots, a_{10}) : count = 20$$

$$(b_1, b_2, a_3, a_4, \dots, a_{10}) : count = 10$$

$$(b_1, a_2, b_3, a_4, \dots, a_{10}) : count = 20$$

$$(a_1, b_2, b_3, a_4, \dots, a_{10}) : count = 50$$

الف) data cube کامل دارای چند aggregate cell می باشد؟

ب) در صورتیکه شرط iceberg به صورت $count \geq 70$ باشد، iceberg cube دارای چه تعداد aggregate cell می باشد؟

ج) data cube کامل دارای چند سلول بسته است؟

سؤال ۳

همانطور که می‌دانید چند روش مختلف برای محاسبات data cube ها وجود دارد مانند:

- Multiway Array Computation

- Bottom-Up Computation (BUC)

الف) این روش‌ها را به صورت مختصر توضیح دهید.

ب) فرض کنید که می‌خواهیم یک data cube کامل چگال اما با تعداد ابعاد پایین را محاسبه کنیم. روش‌های ذکر شده را برای این منظور با یکدیگر مقایسه کنید.

ج) فرض کنید می‌خواهیم یک iceberg cube با تعداد بعدهای پایین را محاسبه کنیم و توزیع داده‌ها بسیار ناهمگون است (چولگی داده‌ها بالاست). روش‌های ذکر شده را برای این منظور با یکدیگر مقایسه کنید.

د) فرض کنید هدف محاسبه‌ی یک iceberg cube بسیار sparse با تعداد بعدهای بسیار بالا است. در این شرایط نیز روش‌های مذکور را با یکدیگر مقایسه کنید.

سؤال ۴

جدول زیر که مربوط به داده‌ی فروش محصولات یک فروشگاه است، را در نظر بگیرید.

Product Type and Model	Date	Price	Count
Epson Printer	Jan 2014	500	20
HP Printer	Jan 2014	500	10
Canon Scanner	Nov. 2013	650	30
Dell Laptop	Jan 2014	1000	10
HP Laptop	May 2013	1000	40
Apple Laptop	May 2013	2000	10
Sony Printer	Nov. 2013	650	20
Canon Scanner	Jan 2014	650	100
Epson Printer	March 2012	1000	10
Canon Scanner	May 2013	650	50
HP Laptop	Jan 2014	1000	10
Epson Printer	Nov. 2013	1000	30

با استفاده از روش Multiway Array Aggregation، data cube کامل را برای این داده‌ها محاسبه کنید.

Aggregate measure = sum(count) می‌باشد و نیازی به chunk بندی نیست.

ملاحظات (حتما مطالعه شود)

گزارش شما باید در قالب یک فایل با فرمت PDF (گزارش تایپ شده) و با عنوان DM_HW2_StudentID تحویل داده شود.

- خوانایی گزارش شما از اهمیت ویژه‌ای برخوردار است. به تمرین‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند، ترتیب اثری داده نخواهد شد.
- مهلت تحویل تمرین به هیچ عنوان تمدید نخواهد شد. تمرین تا یک هفته بعد از مهلت تعیین شده با جریمه تحویل گرفته می‌شود که جریمه تاخیر تحویل تمرین تا یک هفته ۳۰ درصد است.
- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارئه شده باید نتیجه فعالیت فرد نویسنده باشد (همفکری و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب به همه افراد مشارکت کننده، نمره تمرین صفر و به استاد نیز گزارش می‌گردد.
- در صورت بروز هرگونه مشکل با خانم شیما قاسمی از طریق ایمیل زیر در ارتباط باشید:

ghasemi.shima@gmail.com

مهلت تحویل بدون جریمه: ۳۰ فروردین ۱۴۰۱

مهلت تحویل با تاخیر، با جریمه ۳۰ درصد: ۶ اردیبهشت ۱۴۰۱