درس طرحی پایگاه داده

تمرین تئوری سری پنجم ۱۴۰۴/۰۳/۲۱

متین باقری ۴۰۲۱۰۵۷۲۷

---

۱) آ) غلط، داده های خود را مستقیما از operational or transactional data sources می‌گیرند.

ب) صحیح

ج) اگر منظور update شدن دائم و سریع است، خیر (غلط). چرا که هدف آنها نگه داری داده های stable, consistent, historical است. اما در بازه های زمانی طولانی از پیش تعیین شده، می توانند آپدیت شده و یا سطر های جدیدی در آنها load شود. توجه داریم که این آپدیت ها real-time نیست. برای مثال: برخی از رکوردها، مانند آدرس‌های مشتریان، می‌توانند به‌روز شوند در حالی که مقادیر قبلی همچنان حفظ می‌شوند. (صحیح)

د) صحیح

ه) صحیح، هدف اصلی دسترسی به داده های consistent, historical برای تصمیم گیری و تحقیق است.

و) صحیح

ز) غلط، sql مناسب تر است، به علت ساختار یافته بودن داده ها و داشتن عملگر های بیشتر مانند window, join , ...

ح) غلط، noSQL ممکن است real-time consistency نداشته باشد و همچنین برای چنین استفاده های مشخصی استفاده از پایگاه داده ساختار یافته مناسب تر است.

ط) غلط، به معنای ذخیره داده بر روی چندین سرور برای کاهش احتمال از دست رفتن داده و افزایش availability است.

ی) صحیح

ک) صحیح

ل) صحیح

م) غلط، به علت اینکه ممکن است داده های روی سرور ها با هم تطابق نداشته باشند و یا اینکه مجبور باشیم داده های چندین سرور را با هم ترکیب کنیم، ممکن است کند تر باشد.

---

۲) آ)

- data redundancy
- inconsistent data in different data marts
- نداشتن یک پایگاه شامل تمامی داده ها برای بررسی های کلی، و نیاز به ادغام داده های mart های مختلف
- دشواری update به علت نیاز به اعمال آن در چندین بخش
- دشوار و پرهزینه بودن افزودن data mart های جدید و یکپارچه سازی آن ها

از نگاه دیگر:

- measure: داده های عموما عددی که در محاسبات و تحلیل ها استفاده می کنیم
- dimension: توضیحات درمورد ماهیت داده های measure

ب)

Enterprise Data Warehouse (EDW): یک مخزن مرکزی که داده‌های historical را ذخیره می‌کند. برای تحلیل‌های تجاری و تصمیم‌گیری‌های استراتژیک

Data Mart: نسخه کوچک‌تری از Data Warehouse که بر یک بخش خاص از کل دیتابیس تمرکز می‌کند. انواع آن شامل Independent Data Mart (مجزا از Data Warehouse) و Dependent Data Mart (وابسته به Data Warehouse) هستند.

Virtual Data Warehouse: ساختاری که داده‌ها را بدون نیاز به ذخیره‌سازی فیزیکی، از منابع مختلف جمع‌آوری می‌کند و با استفاده از query های پیچیده، امکان تحلیل را فراهم می‌کند. در واقع شامل مجموعه ای از view ها است.

---

۳) آ) آ)

۱. Raw Data: داده‌هایی هستند که مستقیماً از سیستم‌های عملیاتی استخراج شده و هنوز پردازش نشده‌اند.

۲. Summary Data: داده‌هایی هستند که برای گزارش‌گیری و تحلیل خلاصه‌سازی شده‌اند.

۳. Metadata: اطلاعاتی درباره ساختار، منبع، زمان بارگذاری و سایر ویژگی‌های داده‌ها هستند.

ب) Data Mart به بخش خاصی از انبار داده گفته می‌شود که برای یک واحد یا بخش خاص سازمان طراحی شده است.

- افزایش سرعت پاسخگویی به query های آن بخش
- کاهش حجم داده‌ها برای کاربران خاص هر mart
- ساده‌سازی و کنترل دسترسی به داده‌ها برای تحلیلگران، چون داده های هر بخش تفکیک شده اند
- عدم نیاز به یک حافظه بسیار بزرگ

ج) OLAP مخفف Online Analytical Processing است و برای تحلیل داده‌ها به‌صورت چندبعدی استفاده می‌شود. برخلاف OLTP که برای عملیات روزمره کاربرد دارد، OLAP برای تحلیل داده‌های بزرگ کاربرد دارد.

- امکان تحلیل چندبعدی و سریع داده‌ها
- سرعت بالا در اجرای کوئری‌های تحلیلی پیچیده

- امکان استفاده از ابزارهای گرافیکی و داشبوردهای مدیریتی
- سرعت بیشتر در اجرای aggregate functions

د)

ROLAP (Relational OLAP): داده‌ها در پایگاه داده رابطه‌ای ذخیره می‌شوند و از جدول‌های معمولی برای تحلیل استفاده می‌شود.

OLAP (Multidimensional OLAP): داده‌ها در ساختارهای چندبعدی اختصاصی ذخیره می‌شوند که سرعت تحلیل را بالا می‌برد.

HOLAP (Hybrid OLAP): ترکیبی از ROLAP و MOLAP است که مزایای هر دو را با هم دارد، یعنی هم ذخیره‌سازی مؤثر دارد و هم سرعت بالا در تحلیل.

ه)

roll-up: با group by داده ها را به سطح کلی تر برده و مجموع آنها را حساب میکند

dice: روی بیش از یک بعد از داده قید اعمال کرده

slice: قید فقط بر روی یک بعد

و)

- Star Schema: ساده‌ترین مدل که دارای یک جدول مرکزی (Fact Table) است که به جداول جانبی (Dimension Tables) وصل می‌شود. برای پرس‌وجوهای سریع و تحلیل داده‌ها استفاده می‌شود.
- Snowflake Schema: مشابه star است، اما جداول جانبی به جداول کوچکتر تقسیم می‌شوند که می‌توانند به هم لینک شده باشند.

- Galaxy Schema: شامل چندین جدول Fact و جداول Dimension مشترک است. برای تحلیل‌های پیچیده و چند بعدی
- Fact Constellation Schema: انعطاف‌پذیری بیشتری در ارتباطات داده‌ها دارد. برای تحلیل‌های چندبعدی و مدل‌سازی داده‌های سازمانی

---

۴) آ) توزیع بار به پخش کردن درخواست‌های پردازشی بین چندین سرورگفته می‌شود تا کارایی و سرعت سیستم افزایش پیدا کند. وقتی کاربران زیادی به سیستم درخواست ارسال می‌کنند، فشار روی یک سرور کم شده و اطلاعات به‌طور تقریبا برابر در سراسر شبکه مدیریت می‌شوند.

ب) noSQLs نمی‌توانند همزمان هر سه ویژگی زیر را داشته باشند و حداکثر ۲ تا از آنها را دارند:

Consistency: همه‌ی connection ها داده‌های یکسانی را نمایش می‌دهند، یعنی اگر یک کاربر داده‌ای را تغییر دهد، تغییر فوراً در همه‌ی سیستم‌ها اعمال می‌شود.

Availability: سیستم همیشه در دسترس است و درخواست‌ها را بدون توقف پردازش می‌کند، حتی اگر یکی از سرورها دچار مشکل شود.

Partition Tolerance: در صورت قطعی ارتباط بین سرورها، سیستم همچنان قادر به ادامه‌ی کار و پردازش درخواست‌ها خواهد بود (با استفاده از partion های موجود و در دسترس)

ج)

```
db.products.find(
   { "price": { $gt: 1000 } },
   { "Name": 1, _id: 0 }
)
```

```
db.products.aggregate([
    { $group: { _id: "$category", totalStock: { $sum: "$Stock" } } }
 ])
```