درس تحلیل ها و سیستم های داده های حجیم: تمرین سری اول

استاد: آقای دکتر سید حسین خواسته

دستیاران: پژمان زیوری، سمانه تائبی

۱- در مورد مفاهیم column store و row store توضیح دهید و مزایا و معایب هریک را بیان کنید.

۲- frequent items

برای آیتم ست زیر، آیتم های پرتکرار را با در نظر گرفتن min support=3 با استفاده از الگوریتم fp-growth استخراج نمایید. (با رسم شکل برای هر مرحله)

Transaction ID	List of items
T1	B , A , T
T2	A , C
T3	A , S
T4	B , A , C
T5	B , C
T6	A , S, D
T7	B , S
T8	B , A , S , T
T9	B , A , S

۳- کار با پایگاه داده

✓ **هدف از این تمرین اشنایی عملی با دیتابیس های NoSQL به عنوان دیتابیس های مقیاس پذیر اکوسیستم کلان داده است.**

✓ **سه سامانه متداول مدیریت بانک اطلاعاتی به شرح زیر می باشد:**

۱. MongoDB به عنوان نماینده بانک های اطلاعاتی Document Oriented

۲. HBase به عنوان نماینده بانک های اطلاعاتی سطرگسترده

۳. Neo4j به عنوان نماینده بانک های اطلاعاتی گراف محور

✓ **در این تمرین با سامانه مدیریت بانک اطلاعاتی MongoDB به عنوان نماینده بانک های اطلاعاتی سندگرا (Document Oriented) کارخواهیم کرد.**

نکته: دقت کنید که بانک های اطلاعاتی NoSQL را معمولاً به عنوان دیبابیس های جانبی (Not Only SQL) و بسته به نیاز خاصی که در کسب وکار داریم استفاده می کنیم .بنابراین آنها را به عنوان جایگزین های SQL در نظر نگیرید.

سوال ۱: دیتاستی با نام (store.txt) را در نظر بگیرید که شامل تراکنش های یک فروشگاه است.

نام دیتاست	رکورد نمونه	توضیح
store	(U,V,T)	کاربر U کالای V را به قیمت T از فروشگاه خریداری کرده است.

✓ **برای دیتاست یک Collection بسازید به طوری که هر سطر ان به صورت زیر باشد:**

{‘Customer_id’:<value>, ‘goods_id’:<value>, ‘Price’: <value>}

✓ **از پایگاه داده ای که درست کردید Dump بگیرید.**

✓ **Collection را روی Customer_id شاخص(Index) کنید.**

✓ **به پرسش های زیر پاسخ دهید.**

دستورات اصلی

۱- کاربر با ایدی 50 کدام کالا ها را خریداری کرده است.?

۲- کالا با ایدی 3000 توسط چند نفر خریداری شده ?

دستورات تجمعی و آماری (Aggregate Functions)

۱- قصد ارسال هدیه به مشتریانی را داریم که مجموع قیمت خرید ان ها از 90 دلار بیشتر است. ایدی این مشتریان به نحوی می خواهیم که به صورت صعودی مرتب باشد.

۲- هر کاربر چند کالا خریداری کرده است؟

۳- تعداد مشتریان را برای هر کالا را به تفکیک لازم داریم .چگونه این اطلاعات را تولید میکنید ؟

۴- مشتری با بیشترین مجموع خرید و مشتری با کمترین مجموع خرید کدامند؟
۵- میانگین خرید هر مشتری را به تفکیک نیاز داریم؟

بررسی کارآیی شاخص

با توجه به ساختار انعطاف پذیر مانگو، استفاده از شاخص ها درفیلدهایی که درجستجوها، به کرات استفاده می شوند نقش مهمی در کارآیی برنامه ما خواهد داشت . تعیین Customer_id به عنوان شاخص روی زمان اجرای کدام یک از سوالات ۱-۵ اثر دارد؟ میزان افزایش سرعت پاسخگویی به ان سوالات برای زمانی که Customer_id شاخص نیست چه مقدار تفاوت دارد برای یکی از سوالات گزارش کنید.

سوال ۲: دیتاستی با نام (data.csv) را در نظر بگیرید.

۱. لیست از وضعیت کلی برای روز چهارم هفته باشگاه به ترتیب تعداد افراد.
۲. مجموع افرادی که در ماه هشتم به باشگاه می‌ایند.
۳. مجموع افرادی که در تعطیلات به باشگاه می‌ایند.
۴. پنج روز از شلوغ‌ترین روزهای باشگاه به ترتیب.
۵. میانگین دما و وضعیت تعطیلی در روهای خلوت باشگاه به ترتیب تاریخ.
۶. مجموعه افرادی که در روز دوم و ماه دهم از ساعت ۱۷ تا ۲۲ به باشگاه میایند.

نکات تحویل

- در انتهای ترم و بعد از ارسال همه تمرینها، به صورت آنلاین تمرینها تحویل گرفته خواهد شد و یا فقط کسانی که مشکوک به تقلب هستند.
- هدف تمرین یادگیری هر چه بهتر مطالب ارائه شده در کلاس است .لذا از کپی کردن جدا خودداری کنید. تشابه غیرمنطقی بین گزارش ها و کدهای ارسالی تقلب محسوب شده و نمره تمرین تمامی افراد شرکت کننده در آن صفر در نظر گرفته خواهد شد.
- برای پاسخ های خود گزارشی تهیه کنید و به همراه کد ضمیمه کنید. در گزارش خود توضیحی مختصر از کد خود ارائه کنید.
- مشکلات خود را در گروه درسی مربوطه مطرح کنید.
- پاسخ های خود را در یک فایل فشرده به صورت HW1_[Lastname]_[StudentNumber].rar ذخیره کنید.

با آرزوی بهترین ها