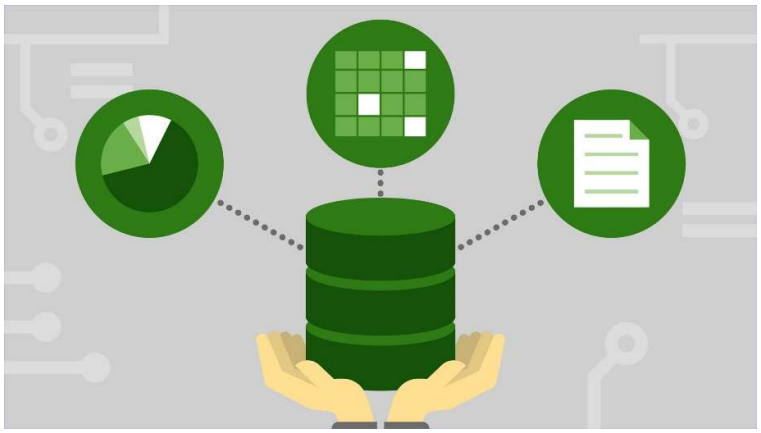
به نام خدا

دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده برق و کامپیوتر



آزمایشگاه پایگاه‌داده

دستورکار شماره ۳

آشنایی با دیتابیس‌های گراف‌محور

مهرماه ۹۹

سید مجتبی بنائی

مقدمه

با توجه به اینکه دو دستورکار قبلی، نیاز به صرف وقت زیادی داشت، تصمیم گرفتم برای اینکه کمی حال و هوای درس عوض شود، آشنایی با دیتابیس Neo4j به عنوان یک دیتابیس گراف‌محور را به جای جلسات پایانی، به این جلسه منتقل کنم.

این دستور کار، بخشی از تمرین‌های درس کلان‌داده ترم زمستان ۹۸ دانشگاه است که دیتابیس و نیازمندیهای اطلاعاتی (سوالات) به شما داده می شود و پاسخ آنها هم ضمیمه این دستور کار است. کافی است بر اساس یکی از این سه پاسخ‌نامه ای که در اختیار شما قرار گرفته است، دیتابیس را نصب، سوالات را اجرا و خروجی را به همراه توضیح هر کدام (یعنی دستور اجرا شده دقیقا چه کاری انجام می دهد)در گزارش نهایی خود بیاورید.

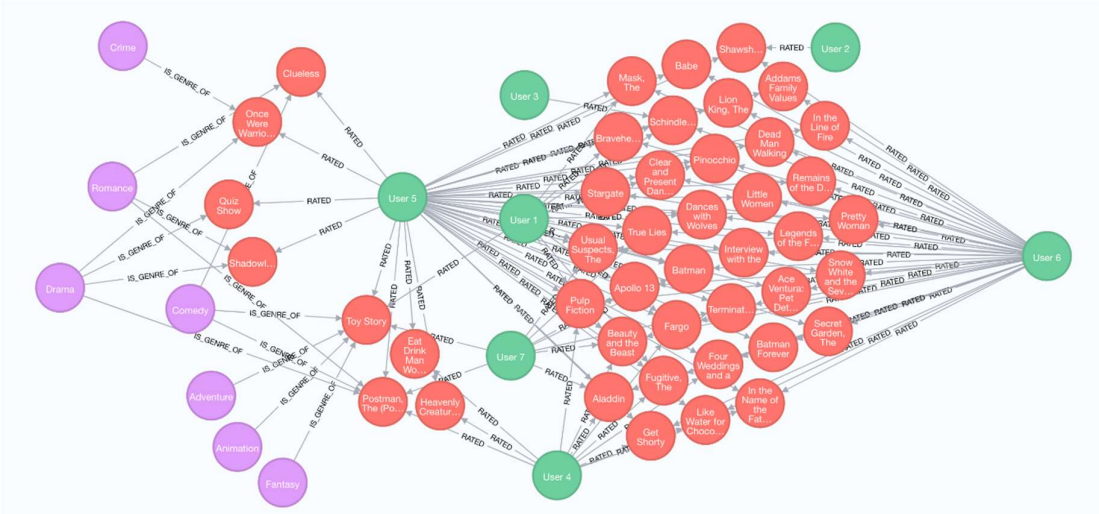
توضیح اینکه در بسیاری از کاربردهای واقعی، ما با موجودیتها و ارتباطات آنها سروکار داریم. این موجودیتها می توانند به عنوان نودهای گراف و ارتباطاتشان به عنوان یالهای گراف در نظر گرفته شوند که هر کدام، می توانند حاوی خصوصیاتی نیز باشند. این موضوع، پایه و اساس دیتابیس‌های گراف محور است. در این نوع از دیتابیس‌ها، به جای کوئری‌های SQL، پیمایش گراف انجام میدهیم.

در این دستور کار ، تنها فایل پی دی اف خروجی برای ارسال ، کافی است و نیازی به ضمیمه کردن کوئری ها نیست. از طرفی، سعی کنید تا جایی که امکان دارد از امکانات نمایش گراف در خروجی های هر سوال، استفاده کنید.

مدلسازی داده‌ها با گراف

در این بخش به بررسی عملی یک بانک‌اطلاعاتی گراف محور رایج و محبوب با نام Neo4j خواهیم پرداخت.

دیتاست مورد استفاده در این بخش، دیتاست معروف Movielens[1] که پایگاه داده ای از فیلم‌ها، بازیگران، کاربران و امتیازات داده شده هر کاربر به هر فیلم است. نمونه‌ای از مدلسازی این دیتابیس با گراف ، با سه نوع نود *فیلم*، *کاربر* و *ژانر* در زیر نمایش داده شده است :



با توجه به اینکه هدف اصلی از این تمرین، کار با دیتابیس گراف‌محور Neo4J است، کدهای پایتون لازم برای مدلسازی و درج اطلاعات در این دیتابیس در اختیار شما گذاشته شده است.

پیش نیازها :

- دانلود دیتاست MovieLens[2] - نسخه small .(که حاوی ۱۰۰ هزار امتیاز و ۹۰۰۰ فیلم است.)
- دانلود Neo4j و نصب آن [3] – می توانید از نسخه داکر آن هم استفاده کنید.
- درج دیتاست دانلود شده در Neo4j از طریق نمونه کد نوشته در این آدرس[4] - پوشه ingestion
- کار با داده‌ها و آشنایی با زبان Cypher به کمک مثال‌های مرتبط با دیتاست تمرین در این آدرس[5]

خلاصه ای از کارهای انجام شده در این بخش را در ابتدای گزارش این بخش حتما ذکر کنید. (در حد یک یا دو پاراگراف) موفق باشید ....

[1] https://grouplens.org/datasets/movielens/
[2] http://files.grouplens.org/datasets/movielens/ml-latest-small.zip
[3] https://neo4j.com/docs/operations-manual/current/installation/windows/
[4] https://github.com/tkcsdvd/neo4j-movielens
[5] http://guides.neo4j.com/sandbox/recommendations

سوالات و پرس و جوهای مورد نیاز

خروجی اصلی این بخش، پاسخ به پرس و جوهای زیر از طریق Neo4j خواهد بود :

1. چه ژانرهایی در این دیتاست وجود دارد ؟
2. تعداد کل فیلم‌ها چقدر است ؟
3. تعداد امتیازات داده شده به فیلم Silence of the Lambs در هر رده امتیازی را به دست آورید (چه تعداد امتیاز ۱، چه تعداد امتیاز ۲ و ....)
4. تعداد امتیازات هر فیلم را به دست آورید و نتایج را به صورت نزولی مرتب کنید.
5. کدام ژانر، بیشترین فیلم تولید شده را به خود اختصاص داده است ؟ کدام ژانر بیشترین امتیاز کاربران و کدام ژانر کمترین امتیاز را به دست آورده است ؟ (میانگین امتیاز برای هر ژانر را در نظر بگیرید - برای انجام این مرحله هم احتمالا نیاز به پیش پردازش خواهید داشت)
6. عنوان فیلم‌های تولید شده در سال ۲۰۰۰ را به دست آورید.
7. با توجه به فیلد Occupation در فایل users.dat ، برنامه نویسان، به کدام فیلم بیشترین امتیاز را داده‌اند؟ (کدام فیلم، تعداد امتیاز ۵ بیشتری دریافت کرده است/ نیاز به میانگین‌گیری نیست)
8. پنج فیلم محبوب رده سنی ۱۸ تا ۳۴ سال، را به دست آورید. (میانگین امتیاز را محاسبه کنید / رده سنی در فایل users.dat قرار گرفته است.)
9. بیست فیلم محبوب را به دست آورید(بالاترین میانگین امتیازات) . با این شرط که حداقل ۳۰ کاربر به آنها رای داده باشند.