- در این پروژه لازم است در مورد پایگاه‌های داده NoSQL اطلاعاتی را جمع‌آوری کرده، سپس با توجه به هدفی که در این پروژه دنبال می‌شود، پایگاه داده مناسب را انتخاب و برای نگهداری نتایج از آن استفاده کنید.
- هدف این پروژه استخراج اطلاعات تعدادی از پروازها از وب‌سایت‌های هواپیمایی، ذخیره‌سازی این اطلاعات در یک پایگاه داده مناسب و انجام پردازش‌هایی روی آنها است.
- زبان برنامه‌نویسی مورد استفاده دلخواه است.
- برای ارتباط با پایگاه داده، ذخیره و دریافت نتایج نمی‌توانید از ORM/OGM/ODM/... استفاده کنید، بلکه باید با استفاده از واسط‌هایی همچون JDBC/ODBC/ADO.NET به پایگاه داده متصل شده و پرسمان‌های لازم را بنویسید.

توضیحات

این پروژه دارای دو مرحله است:

- در مرحله اول باید اطلاعات را از چند وب‌سایت هواپیمایی با کراول کردن[1] (خزش وب)، استخراج کرده و در پایگاه داده ذخیره کنید.
- در مرحله دوم باید ضمن آماده کردن پرسمان‌ها، این امکان را در برنامه خود به وجود آورید که بتوان اطلاعاتی را از پایگاه داده استخراج کرد.

[1] Crawling

مراحل پروژه

۱. مرحله اول – استخراج اطلاعات و قرار دادن در پایگاه داده

در ابتدای کار باید اطلاعات پروازهای خارجی را که در یک هفته آینده از ایران خارج و یا به ایران وارد می‌شوند، از وب‌سایت‌هایی همچون وب‌سایت ۱، وب‌سایت ۲، وب‌سایت ۳ و یا وب‌سایت‌های جهانی مشابه از هواپیمایی‌های مختلف استخراج کنید.

- ✓ توجه کنید که باید حداقل دو وب‌سایت را کراول کنید و اطلاعات آنها را در پایگاه داده ذخیره کنید.
- ✓ ممکن است برای برخی پروازها اطلاعات اضافه‌ای در سایت دیگر وجود داشته باشد، در این صورت این اطلاعات نیز باید در پایگاه داده ذخیره شوند. به عنوان مثال، ممکن است در سایتی برای یک پرواز، اطلاعاتی همچون امکانات پرواز، غذا، محدودیت سنی، ویزا و ... ذکر شده باشد. واضح است که ذخیره چنین اطلاعاتی در پایگاه‌های داده رابطه‌ای با مشکلاتی مانند زیاد شدن خانه‌های خالی، اجبار به تغییر یا افزودن جدول‌ها برای اطلاعات جدیدتر همراه است.

نحوه استخراج اطلاعات:

برای استخراج اطلاعات می‌توانید بطور دلخواه از کتابخانه‌های آماده برای کراول کردن استفاده کنید. برای مثال در زبان جاوا می‌توانید از کتابخانه‌های JSoup یا Crawler4J و یا از برخی ابزارهای سطح پایین‌تر مثل Selenium استفاده کنید. نمونه‌ای ساده از JSoup را می‌توانید در اینجا ببینید.

- ✓ به طور کلی، در عملیات کراول کردن درخواست‌هایی به سمت یک وب‌سایت ارسال شده و نتایج در قالب تعدادی تگ html بازگردانده می‌شوند. سپس، پردازش‌های لازم بر روی این نتایج اعمال شده و قسمت‌های مورد نظر استخراج می‌گردد. (توضیحات بیشتر در جلسه آخر حل تمرین عملی مورخ ۱۴۰۰/۱۰/۱۱)

3

- ✓ همچنین باید توأم با ذخیره اطلاعات کراول شده در پایگاه داده، زمان انجام کراول (*time step*) نیز ثبت گردد. به این ترتیب، در صورت کراول مجدد، می‌توان آخرین داده‌ها را از روی زمان کراول آنها تشخیص داد.

۲. مرحله دوم – پرسمان‌ها و قابلیت‌های برنامه پس از استخراج اطلاعات از وب‌سایت‌ها

پس از کراول کردن و انتقال اطلاعات به پایگاه داده NoSQL، در این مرحله باید پرسمان‌های مناسب را برای انجام کارهای زیر نوشته و امکان اجرای آنها را از طریق کنسول (رابط کاربری ساده) برای کاربر فراهم آورید.

- ✓ توجه کنید اگر نیاز است اطلاعات خاصی به کاربر نشان داده شود، باید فیلتر شدن اطلاعات و یا مواردی مانند به دست آوردن مجموع یا میانگین در بدنه پرسمان انجام شود و نوشتن کد برای فیلتر کردن مجاز نیست.

قابلیت های برنامه:

۱. امکان مشاهده همه پروازها در یک روز مشخص.
۲. امکان مشاهده همه پروازها در یک بازه قیمتی معین (به یورو یا دلار - نکته چهارم را بخوانید!)
۳. امکان مشاهده حداقل و حداکثر قیمت پروازها از مبدأ به مقصد مشخص.
۴. امکان مشاهده میانگین و مجموع قیمت پروازها از مبدأ به مقصد مشخص.
۵. امکان مشاهده موارد ۱ تا ۴، در صورتی که نوع پرواز مشخص باشد. (مثلاً First Class)
۶. امکان مشاهده پروازها با مبدأ و مقصد مشخص در یک بازه قیمتی و تعیین ارزان‌ترین پرواز.
۷. امکان مشاهده پروازها با مبدأ، مقصد و ظرفیت معین.
۸. موارد ۶ و ۷ به همراه بازه زمانی.
۹. امکان مشاهده شرکت‌های هواپیمایی موجود برای پروازها از یک مبدأ به مقصد مشخص در تاریخ معین.
۱۰. حذف موارد بر اساس تاریخ و نام شرکت هواپیمایی.

۱۱. امکان تغییر ظرفیت یک پرواز مشخص.

۱۲. امکان تغییر ظرفیت همه پروازهای یک شرکت هواپیمایی در تاریخ، مبدأ و مقصد مشخص.

۱۳. امکان مشاهده نام فرودگاه‌های مبدأ و مقصد موجود برای پروازهایی با یک مبدأ و مقصد (کشور) و تاریخ مشخص.

۱۴. از موارد بالا، همه اطلاعاتی که به صورت تجمعی نیستند (یعنی خروجی یک لیست است) باید به صورت صفحه‌بندی شده با تعداد دلخواه توسط کاربر قابل مشاهده باشند.

۱۵. موارد بالا را باید بتوان هم به صورت صعودی و هم نزولی بر اساس قیمت و یا تاریخ مشاهده کرد. (بجز موارد ۱۰ تا ۱۳)

راهنمایی انتخاب پایگاه داده

برای انتخاب پایگاه داده به موارد زیر توجه نمایید:

- ✓ این پروژه با حجم زیادی از داده‌ها که فاقد ساختاری مشخص است، سر و کار دارد.
- ✓ واضح است که ساختار داده‌ها به صورت گرافی، سری زمانی و key-value نیست و استفاده از این مدل‌ها در این پروژه کارایی ندارد.
- ✓ به نظر می‌رسد مدل‌های document-based یا column-family برای این پروژه انتخاب مناسبی باشند. بنابرین، انتظار می‌رود در مورد تفاوت‌های پایگاه‌های داده یاد شده جستجو کرده و مناسب‌ترین گزینه را انتخاب کنید.

نکات مهم

۱. برنامه می‌تواند در محیط کنسول اجرا شود و لزومی بر ایجاد محیط گرافیکی (GUI) نیست.

✓ ایجاد محیط گرافیگی مناسب دارای **امتیاز تا سقف ۰,۵ نمره** خواهد بود.

۲. استفاده از ORM امتیاز محسوب نمی‌شود!

۳. استفاده از API (**Application Programming Interface**) مشروط بر کراول کردن حداقل یک وب‌سایت دارای **امتیاز تا سقف ۰,۵ نمره** خواهد بود.

۴. توجه کنید قیمت پروازها ممکن است در واحدهای ارزی مختلفی باشند؛ در این صورت قبل از کراول کردن، قیمت ارزهای مختلف و نسبت آنها را از یکی از وب‌سایت‌های ارزی (مانند iranjib و tjgu) استخراج کرده و تمام قیمت‌ها را به دلار یا یورو تبدیل و در پایگاه داده ذخیره کنید.

۵. اگرچه نصب پایگاه داده ارجح است، اما لزومی بر نصب فیزیکی آن بطور کامل نیست و می‌توانید از docker image پایگاه داده مورد نظر استفاده کنید. توجه کنید جهت از بین نرفتن داده، حتماً از volume استفاده نمایید.

❖ تحویل این پروژه نیز در دو مرحله بارگذاری در سایت و تحویل شفاهی (آنلاین) بصورت گروهی است.

❖ زمان تحویل شفاهی در بازه ۱۱ تا ۱۳ بهمن خواهد بود که متعاقباً در کانال تلگرام درس اعلام می‌شود.

❖ حضور تمام اعضای گروه در ارائه شفاهی الزامی است.

❖ در صورت عدم ارائه شفاهی هیچ نمره‌ای به این فاز تعلق نخواهد گرفت.

❖ سؤالات خود را می‌توانید با دستیاران آموزشی درس (*@Shayan_Daneshvar* و *@Ali_E99*) مطرح نمایید.