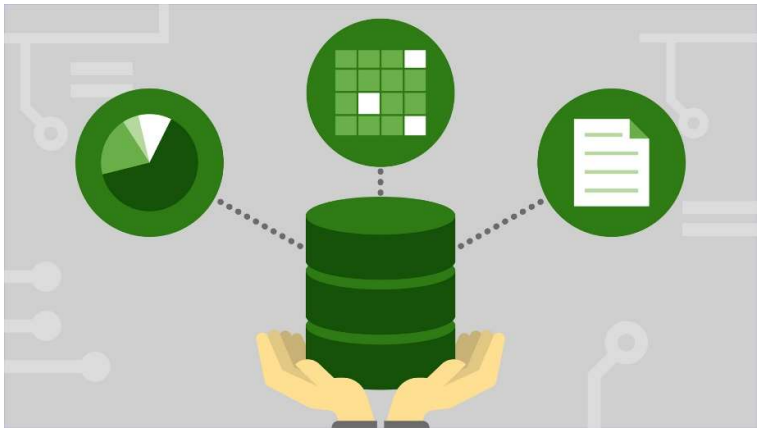
به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر



آزمایشگاه پایگاه‌داده

دستورکار شماره ۱۰

مهلت تحویل :

۱۴۰۱/۰۳/۲۰

مجتبی بنائی

دستور کار شماره ۱۰ - کار با آپاچی درواید و کافکا

یکی از دیتابیس‌هایی که اخیرا و به صورت روزانه با آن سروکار دارم و به کارگیری آنرا در بسیاری از شرکتهای متوسط و بزرگ ایران برای ذخیره و پاسخگویی به حجم عظیم داده‌های ورودی ، یک ضرورت میدانم ، دیتابیس تحلیلی آپاچی درواید است.

این دیتابیس که در رسته بانک‌های اطلاعاتی تحلیلی به عنوان نسل جدید دیتاویرهوس‌ها قرار میگیرد، ویژگی های کاربردی بسیار خوبی دارد. به عنوان نمونه :

ـ یک دیتابیس تحلیلی است و ذخیره داده‌ها به صورت ستونی در آن، هم حجم داده‌ها را بسیار کم میکند و هم سرعت پاسخگویی به کوئری های تحلیلی که معمولا ترکیبی از گروه‌بندی و فیلترینگ داده‌ها هستند را بسیار بهبود می بخشد.

ـ معماری توزیع شده بسیار پیشرفته ای دارد که مقیاس پذیری و پاسخ گویی به هر حجمی از داده‌ها در زمان مناسب (زیر یک ثانیه) تضمین میکند (البته نیاز به تخصیص منابع و پایش دارد)

ـ یک دیتابیس سری زمانی است و تمامی داده‌ها باید حاوی مهر زمان یا تایم استمپ باشند.

ـ بخش دریافت داده یا اینجستشن بسیار پیشرفته ای دارد و کافی است داده‌های خود را به Kafka|MinIo|HDFS|... منتقل کنید و ادامه کار یعنی دریافت از کافکا ـ و سایر منابع ـ ، فیلترینگ و پردازش اولیه داده ها و ذخیره آنها به صورت بلادرنگ را به درواید بسپارید.

ـ امکان تجمیع داده ها به صورت خودکار در آن فراهم شده است یعنی مثلا داده های هر پنج دقیقه را تجمیع کرده ، آمار مورد نیاز آن بازه زمانی را به صورت خلاصه برای شما محاسبه میکند.

ـ الگوریتم های تقریبی پیشرفته ای دارد که همزمان با تجمیع داده‌ها میتوانید با تقریب بسیار مناسبی، آماره هایی راجع به داده هایی که در حین تجمیع، حذف می شوند را ذخیره کنید.

با این مقدمه، دستور کار آخر را برای آشنایی با این دیتابیس خوش آتیه و نیز بروکر (توزیع کننده پیام در یک شبکه) کافکا در نظر گرفته ام.

توضیح اینکه امروزه کافکا نقش محوری در توسعه سامانه های مقیاس پذیر اطلاعاتی دارد و نقش واسط بین سامانه های مختلف و نیز انتقال توزیع شده اطلاعات در یک شبکه را بر عهده داردو در این دستور کار هم ، داده هایی را به صورت تصادفی تولید و در یک کانال کافکا (هر پیام در یک گروه قرار میگیرد که به آن، تاپیک میگوییم) قرار میدهیم و به کمک موتور اینجستشن درواید، آنها را به صورت مداوم دریافت و در دیتاسورس های متناظر درواید ذخیره می کنیم .

آموزش یک ساعته ای را برای این دیتابیس و نحوه دریافت اطلاعات از کافکا برای مجموعه آموزشی نیک آموز در یک وبینار یک ساعته در انتهای آذر ماه برگزار کرده ام . فایل ویدئوی این کارگاه عملی را از آدرس زیر دریافت کنید :

https://dl6.nikamooz.com/webcast1400/ApachDruid.rar

نکته : حجم ویدئوها، حدود ۱.۴ گیگابایت است و ممکن است دانلود آنها برای شما هزینه بر باشد . می توانید ویدئوهای فوق را در سه قسمت در کانال آپارات مهندس داده[1] هم به صورت آنلاین مشاهده کنید (قسمت اول و دوم برای این گزارش کار کافی است) و کیفیت پخش آنرا به دلخواه تنظیم کنید که حجم اینترنت بسیار کمتری برای این موضوع، مصرف شود.

[1] https://www.aparat.com/playlist/1265528

فایلهای کارگاه عملی هم که برای این دستور کار به آنها نیاز خواهید داشت در کنار دستور کار، در قالب یک فایل زیپ بارگزاری شده است اما میتوانید آنها را از آدرس زیر (پوشه druid) هم دانلود کنید :

https://gitlab.com/nikamooz_bigdata/webinars

بعد از مشاهده این ویدئو و آشنایی با نحوه نصب و راه اندازی درواید و همینطور کلاستر کافکا (کلاستر : مجموعه نودهایی که با هم، یک هدف واحد را در شبکه بر عهده دارند)، برای دستور کار شماره ۱۰ که ترکیب درواید و کافکا خواهد بود، موارد زیر را انجام و گزارش آنرا آپلود کنید :

- درواید را دانلود و راه اندازی کنید.
- کلاستر کافکا را به کمک داکر راه اندازی نمایید (در خط فرمان، درون پوشه داکر، دستور docker-compose up را اجرا کنید)
- برای اینکه بتوانید بین سیستم خودتان و کافکایی که در داکر راه اندازی شده است، ارتباط برقرار کنید و درواید هم بتواند به آن متصل شود نیاز دارید که سطر

127.0.0.1 kafka

را به فایل hosts ویندوز یا لینوکس اضافه کنید. (kafka نام بروکر یا نود کافکایی است که درون فایل داکر کامپوز تعریف شده است و اگر در فایل داکر کامپوز، چندین بروکر کافکا داشته باشید نام تمام آنها را در این جا باید وارد کنید - هر نام در یک سطر)

یک راهنمای ویرایش فایل hosts ویندوز را در این آدرس می توانید بیابید :

https://howtotech.ir/what-is-windows-hosts-file-how-to-change-hosts-file

- چک کنید که واسط کاربری کافکا یعنی AKHQ در آدرس localhost:8000 قابل مشاهده باشد. تاپیک events_topic با ۳ پارتیشن را در کافکا ایجاد نمایید.

- فایل user_event_producer.py موجود در پوشه codes را اجرا کنید به گونه ای که درون واسط کاربری AKHQ هر نیم ثانیه یک بار، پیامی را مشاهده کنید (می توانید تاپیک را انتخاب کنید و از بخش Live tail آخرین پیام ها را ببینید)

- درون درواید، طبق آموزش ویدئویی فوق، یک اینجسشن برای دریافت داده های موجود در تاپیک events_topic کافکا ایجاد کنید به گونه ای که در پایان کار، دیتاسورس مربوطه ایجاد شود.

- در مرحله آخر، یک Rollup یک ساعته و یک Rollup یک روزه با توجه به دیتای تولید شده ایجاد کنید. منظور از رولاپ، تجمیع اطلاعات در یک بازه زمانی است. برای این منظور، مجددا باید یک اینجسشن بر روی همین تاپیک ایجاد کنید وسپس در بخش configure schema گزینه rollup را علامت بزنید و سپس فیلدهایی که داده ها بر اساس آنها در آن بازه زمانی باید گروه بندی شوند را مشخص کنید (دایمنشن یا ابعاد) و سپس متریک های مورد نیاز مانند مجموع یا تعداد یا ماکزیمم و می نیمم برخی فیلدها را تعیین کنید .با اینکار، به محض دریافت داده ها، بسته به متریک تعیین شده، داده ها تجمیع می شود. با اینکار جدولی ایجاد می کنیم که خلاصه داده های یک ساعت یا یک روز در آن ذخیره شده است و نیاز به کوئری گرفتن از جداول اصلی نداریم . از طرفی می توانیم فقط این جداول تجمیعی را نگهداری کنیم و خود داده های اصلی را حذف و یا به یک دریاچه داده منتقل کنیم. برای تجمیع یکساعته ، query granularity را برابر یک ساعت و برای تجمیع یک روزه ،آنرا برابر روز قرار دهید . با اینکار فیلد __time براساس این مقدار ، در جدول نهایی رند خواهد شد.

با ایجاد یک تجمیع یکروزه ، تمام داده های آنروز در یک رکورد خلاصه شده و در دیتابیس ذخیره خواهند شد. بنابراین ، جدول تجمیعی روزانه باید کمترین میزان رکوردها را داشته باشد و دیتاسورس غیر تجمیعی، بیشترین تعداد رکورد را به خود اختصاص دهد.

- پنج کوئری مختلف بر روی این سه دیتاسورس اجرا و نتیجه را نمایش دهید.کوئری های شما حتما شامل گروه بندی ، بازه زمانی ، تعیین ترتیب ویک شرط عادی باشند.

از مراحل فوق و نیز کوئری ها و نتایج آنها، عکس گرفته و در گزارش کار آخر آنها را ذکر کنید . (در حدی که بیان کننده کار شماباشد)