کلان داده و تحلیل داده‌های حجیم

دکتر محمدعلی نعمت بخش

نیمسال دوم سال تحصیلی

۱۴۰۱-۱۴۰۰

پروژه پایانی

طراحی یک سامانه بلادرنگ برای تحلیل لحظه‌ای، مصورسازی، پیشبینی،
ذخیره‌سازی داده‌های تاکسی اینترنتی
(Elasticsearch, Kafka or Filink, Cassandra, Spark, Redis)

مهلت تحویل : ۲۰ تیرماه ۱۴۰۱

مقدمه

هدف از انجام پروژه نهایی درس کلان‌داده، آشنایی عملی با طراحی یک سامانه کاربردی پردازش داده بلادرنگ و مقیاس‌پذیر با استفاده از ابزار و کتابخانه‌های روز دنیا در حوزه بیگ دیتا است. انتظار میرود پس از انجام این پروژه دیدی تجربی و شهودی نسبت به مفاهیم زیر پیدا کنید:

- صف‌های توزیع شده و نقش محوری آنها در سامانه‌های نوین اطلاعاتی.
- الستیک سرچ و قدرت و کارآیی فوقالعاده آن در مدیریت داده‌ها.
- کاساندرا به عنوان یک دیتابیس سطرگسترده مقیاس‌پذیر کارآمد .
- اسپارک و سهولت پیاده‌سازی الگوریتم‌های پیچیده یادگیری ماشین بر روی حجم عظیم داده به کمک آن.
- کاربا داده‌های سری زمانی.

جزییات پروژه و مستندات مورد نیاز برای هر قسمت، در ادامه آمده است.

چشمانداز کلی سامانه

تاکسی‌های اینترنتی نیاز دارند تا حجم زیادی از درخواست‌ها را در بستر اینترنت منتقل کنند. برای این منظور نیاز است تا ابزارهای مقایس‌پذیری استفاده شود تا پاسخگوی این حجم داده باشد. در این پروژه داده‌ی حجیم تاکسی اینترنتی در اختیار شما قرار گرفته است. برای این منظور نیاز داریم تا سامانه‌ای بلادرنگ برای:

۱. ارسال داده‌ها به صورت استریم و دریافت آن.
۲. ارسال داده‌ها بر روی بستر اسپارک و خوشه بندی داده‌ها به صورت آنلاین
۳. ارسال و ذخیره داده بر روی الستیک‌سرچ و نمایش اطلاعات آماری با استفاده از کیبانا.
٤. ذخیره‌داده‌ها بر روی کاساندرا و تحلیل آن.
٥. ذخیره داده‌ها بر روی ردیس و تحلیل لحظه‌ای داده.
٦. پیش‌بینی تعداد سفرهای ممکن در بازه‌زمانی مختلف با کمک سری زمانی.

روند کلی پردازش داده در سامانه نهایی از قرار زیر خواهد بود:

- **گام اول:** داده‌ها با استفاده از یک API وارد کانال اولیه در کافکا یا فیلینک می‌شود. (انتخاب بستر استریم بر عهده دانشجو می‌باشد) هماهنگی کل پروژه برای گام‌های مختلف همانند دنیای واقعی از این طریق صورت می‌گیرد. (داده‌ها در هر مرحله با یک تاخیر مشخص به کمک این بستر انتقال پیدا می‌کند و به مرحله‌ی بعدی می‌رود)

- **گام دوم:** ابتدا داده‌ها را شافل (فقط برای این مرحله) کرده و با ۲۰٪ داده‌ها عملیات خوشه‌بندی به صورت آفلاین انجام شود و مناسب‌ترین تعداد خوشه‌ی را **با ذکر دلیل** مشخص کنید.

- **گام سوم:** پس از گام ۲، **در همان زمان** که داده‌ها از کانال استیریمی دریافت می‌شود، با کمک اسپارک به صورت **بلادرنگ** عملیات خوشه‌بندی را انجام دهید.

- **گام چهارم:** پس از دریافت داده‌ها از گام سوم، داده‌ها را بر روی الستیک ذخیره شده و برای تحلیل‌های بعدی استفاده می‌شود.

- **گام پنجم:** پس از دریافت داده‌ها از گام چهارم، داده‌ها را بر اساس مختصات خوشه، مختصات، تاریخ برای بازیابی‌های آینده در کاساندرا ذخیره کنید. هدف از این عملیات برای بازیابی سریع زمان‌ها و موقعیت مکانی برای مواقع اضطراری می‌باشد.

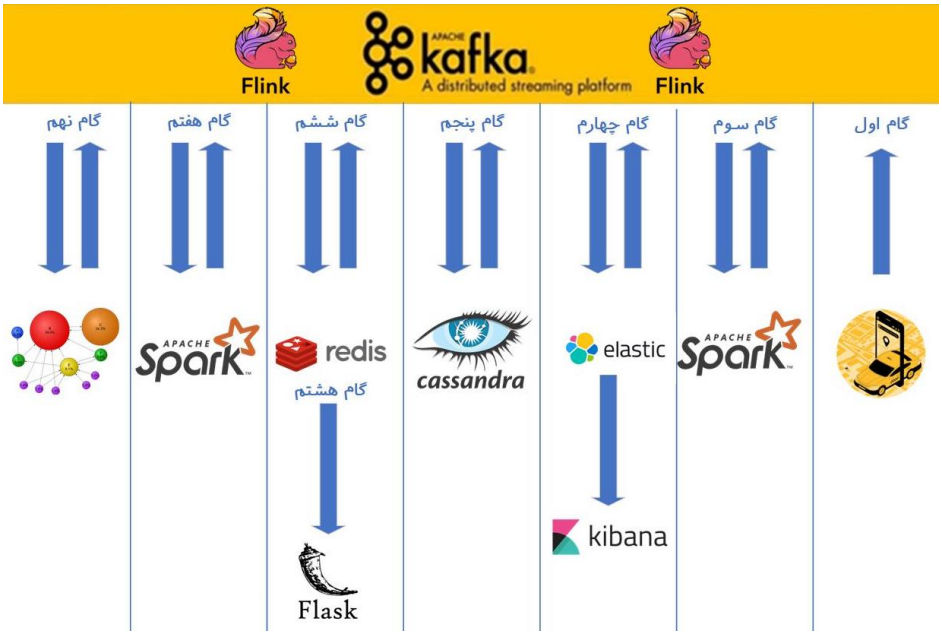
- **گام ششم:** پس از دریافت داده‌های از گام پنجم، اطلاعات آماری و داده‌های لازم برای هر خوشه که در ادامه‌ی سند خواسته شده است در ردیس ذخیره می‌شود.

- **گام هفتم:** با استفاده از اسپارک و الگوریتم رگرسیون تعداد سفرها را برای تایم‌فریم یک‌هفته و یک‌ماه آینده به‌دست آورید.

- **گام هشتم:** طراحی یک وب اپلیکیشن برای نمایش اطلاعات گام ششم که آمار لحظه‌ای داده‌ها توسط یک وب اپلیکیشن و با خواندن داده‌ها از ردیس، به کاربر نمایش داده شود.

- **گام نهم:** داده‌ها را به دو قسمت مساوی تقسیم کنید و مختصات داده‌های دسته‌ی دوم را مقابل دسته‌ی اول قرار دهید. با این کار مختصات اولی نقطه‌ی شروع سفر و مختصات دوم نقطه‌ی پایان سفر را مشخص می‌کند. تایم استپ مجموعه‌ی داده اولی به عنوان زمان داده‌های جدید قرار بگیرد. سپس از با استفاده از اسپارک و الگوریتم پیج‌رنک نقاط مهم و کلیدی را استخراج کنید **(اختیاری)**

شکل ۱ زیر شماتیک معماری این سیستم را نمایش میدهد که محوریت کافکا و نحوه تعامل بخش‌های مختلف آن به خوبی در آن قابل مشاهده است.



شکل ۱- شمای کلی از سامانه بلادرنگ

گام اول

- از بین دو بستر کافکا و فلینک یکی را انتخاب کنید و شروع به دریافت اطلاعات کنید.
- داده‌های دریافتی را با یک شناسه‌ی یکتا (UUID) مشخص کنید.
- تایم استپ دریافت داده‌ها همان ستون Date/Time در مجموعه داده می‌باشد.

گام دوم

ابتدا داده‌ها را شافل (فقط برای این مرحله) کرده و با ۲۰٪ داده‌ها را عملیات خوشه‌بندی به صورت آفلاین انجام شود و مناسب‌ترین تعداد خوشه‌ی را **با ذکر دلیل** مشخص کنید.

گام سوم

- داده‌های دریافتی از استریم به صورت آنلاین و با کمک کتابخانه‌ی اسپارک و تعداد کلاستری که در گام دوم بدست آورده‌اید، خوشه بندی کنید. (کلاسترینگ آنلاین اسپارک)

گام چهارم

- در این مرحله، داده‌های دریافت شده مرحله قبل در الستیک سرچ ذخیره می‌شوند.
- داشبوردی در کیبانا طراحی کنید که موارد زیر را بتوان در آن مشاهده کرد.
    - تراکم نقاط شروع پرتردد در یک بازه‌ی زمانی خاص.
    - ۱۰۰ نقطه‌ی آخر دریافتی از استریم را نمایش دهد.
    - تعداد درخواست‌ها برای ۱۰ نقطه‌ی پر تردد.
    - ۱۰ نقطه‌ی پر تردد در یک بازه‌ی زمانی.

لازم به ذکر است که برای هریک از این کوئری‌ها باید یک نمودار نمایش داده شود.

گام پنجم

- در این مرحله، میخواهیم به کمک کاساندرا و مکانیزم ذخیره‌سازی سطرگسترده آن، تاریخچه زمانی هر نقطه‌ی مختصاتی و زمان سفر(یک هفته‌ای) ذخیره کنیم.
- اگر کاربر نیاز داشت سفرهای اخیر یک نقطه‌ی خاص و یا زمان خاص را ببیند، کافی است داده‌ها از این دو جدول کاساندرا، خوانده شده و به کاربر نمایش داده شود.
- توجه‌شود که ذخیره‌سازی در کاساندرا به صورت مرتب انجام می‌شود و قابلیت جوین بین جداول وجود ندارد.
- دقت کنید که در کاساندرا، تکرار داده‌ها یک اصل کاملا پذیرفته شده است و به دنبال نرمالسازی نباشید.
- جدول‌هایی که باید طراحی شود:
    - جدولی بر اساس کلید زمانی هفته‌ای طراحی کنید.
    - جدولی بر اساس مختصات نقطه‌ی شروع طراحی کنید.
    - UUID رو به عنوان کلید و مقادیر دیگر ویژگی‌ها را در Value ذخیره کنید تا در صورت لزوم بتوان از آن برای بازیابی از الستیک کمک گرفت.
    - جدولی بر اساس زمانبندی ۱۲ ساعته سفرها طراحی کنید.

نکته : تمام این اطلاعات را الستیک سرچ هم می‌تواند با سرعت بسیار بالا در اختیار ما قرار دهد اما هدف از این بخش، آشنایی عملی با کاساندرا و جدا کردن بخشهای مختلف منطقی سامانه از یکدیگر است.

- **کوئری‌ها:**
  - بازیابی سفرهای ۱ روز اخیر، ۶ ساعت اخیر
  - بازیابی سفرها برای یک مختصات خاص
  - بازیابی زمان‌های در خواست برای یک مختصات در تایم یک هفته

گام ششم

به‌ازای هر زمان یک کلید در ردیس در نظر می‌گیریم و با دریافت یک کلید جدید مقدار آن را با یک جمع می‌کنیم. اما چون مثلا بعد از گذشتن یک روز یا یک ساعت، رکوردهای قدیمی باید از آمار فعلی کسر شوند ،بنابراین درطراحی کلیدهای ردیس دقت به خرج دهید. به ازای هر رکورد جدیدی که دریافت می‌کنید، چندین کلید را در ردیس باید به روزرسانی کنید.

راهنمایی : کلیدهایتان را به روز و ساعت مرتبط کنید و با آغاز هر ساعت جدید/ هر روز جدید، کلید جدیدی در نظر بگیرید.

در این مرحله باید بتوانید به سوالات زیر به کمک ردیس که یک دیتابیس مقیم در حافظه بسیار سریع است جواب دهید:

- تعداد سفرها در یک نقطه‌ی خاص در شش ساعت گذشته.
- تعداد کل سفرهای دریافت شده در یک بازه زمانی مثلا روز گذشته.
- تعداد سفرهای دریافت شده در یک ساعت گذشته.
- آخرین درخواست‌های سفر دریافت شده. (یک لیست هزارتایی که با ورود داده‌های جدید، قدیمی‌ها حذف می‌شوند)

دقت کنید که تمام داده‌ها تا یک هفته گذشته باید در حافظه باشند و بعد از آن، باید به صورت خودکار توسط ردیس از حافظه حذف شوند.

با توجه به گام هشتم یک وب اپلیکیشن با فلسک بنویسید که اطلاعات خواسته شده فوق را بتوان درون آن مشاهده کرد. با رفرش کردن صفحه در این اپلیکیشن، آمار آن باید به روز شود.

گام هفتم

با اتصال اسپارک به کاساندرا وخواندن داده‌های ذخیره شده، مدلی پیشبینی کننده‌ای برای تعداد سفرهایی که در آینده انجام می‌شود آموزش داده و ارزیابی کنید.

پیشبینی تعداد سفرهای آینده برای **یک هفته، دوازده ساعت** و **یک ماه** آینده. (شایسته است بخشی از آخرین داده‌ها بر اساس زمان را جدا کرده و آن را پیش‌بینی کنید، پس از آن در یک نمودار مقادیر پیش‌بینی شده از سمت مدل و مقادیر واقعی آن نمایش داده شود.)

گام هشتم

در این مرحله شما تمامی کوئری‌ها، اطلاعات آماری در مرحله ششم با استفاده از یک API در اختیار دیگران قرار دهید. این API باید قابلیت وارد کردن اطلاعات (مثلا بازه‌ی زمانی دلخواه) و دریافت اطلاعات آن به کاربر نمایش دهد.

گام نهم

توضیح : انجام این بخش دارای امتیاز اضافه خواهد بود و انجام آن، اختیاری خواهد بود.

همانطور که پیش‌تر ذکر شد، داده‌ها را به دو قسمت مساوی تقسیم کنید و مختصات داده‌های دسته‌ی دوم را مقابل دسته‌ی اول قرار دهید. با این کار مختصات اولی نقطه‌ی شروع سفر و مختصات دوم نقطه‌ی پایان سفر را مشخص می‌کند. تایم استپ مجموعه‌ی داده اولی به عنوان زمان داده‌های جدید قرار بگیرد. سپس از با استفاده از اسپارک و الگوریتم پیج‌رنک نقاط مهم و کلیدی را استخراج کنید (اختیاری)

نکات قابل توجه

- برای هر گام از پروژه، با یک نرم‌افزار/دیتابیس کار خواهید که بهتر است آخرین نسخه آنها را استفاده کنید .
- شالوده ارتباطی این سامانه، صف توزیع شده کافکا یا فیلینک خواهد بود.
- تعداد اعضای هر تیم، سه نفر است. بهتر است برای هماهنگی بیشتر، یک نفر را به عنوان مدیر تیم انتخاب کرده، هماهنگی و توزیع کارها را انجام دهید.
- داده‌ها از این لینک در دسترس هستند (با توجه به محدودیت درخواست به گوگل حتما داده‌ها را از پیش دانلود کنید)

نکات تحویل

- مهلت ارسال تا ۲۰ تیر ماه خواهد بود.
- انجام این تمرین به صورت تیمی (۳ نفره) می‌باشد و اعضای گروه می‌بایست در صورت سوال به یکدیگرکمک کنند.
- زمان و نحوه‌ی تحویل پروژه به اطلاع شما خواهد رسید.
- هر فرد از اعضای تیم، گزارش آماده شده برای بخش خودش را ارسال خواهد کرد، تا در صورت کم کاری یکی از اعضای تیم، فقط نمره آن فرد، تحت تأثیر قرار گیرد و نمره نهایی، براساس میزان تلاش و مشارکت هر عضو مستقل از بقیه تیم، داده شود. در جلسه تحویل، هر نفر از اعضای تیم به صورت جداگانه کار انجام شده توسط خودش و گزارش آماده شده را تشریح کرده و تسک‌های انجام شده را توضیح خواهد داد. سپس با اجرای پروژه به صورت لوکال و به اشتراک گذاری صفحه نمایش، خروجی واقعی بخش مرتبط با خود را نمایش خواهد داد.
- گزارش شما در فرآیند تصحیح از اهمیت ویژه‌ای برخوردار است، لطفا تمامی مواردی که در تمرین از شما خواسته شده را در گزارش ذکر نمائید.
- کدهای پروژه و به همراه سند به صورت صحیح بر روی گیت آپلود کرده و لینک پروژه را در سند خود قرار دهید. (برای هر فردی سندی با نام خود که شامل فعالیت‌های انجام شده توسط فرد است وجود داشته باشد.)
- تمامی موارد بیان شده (کد و اسناد) در قالب یک فایل zip. در سامانه کوئرا **توسط مدیر گروه** بارگذاری شود.

Project_[Leader_Last_Name]_[StudentNumber].zip