درس

کلان داده و تحلیل داده‌های حجیم

دکتر محمدعلی نعمت بخش

نیمسال دوم سال تحصیلی ۱۳۹۹- ۱۴۰۰

پروژه پایانی

طراحی یک سامانه بلادرنگ برای تحلیل لحظه‌ای داده‌های پیام‌رسان‌های داخلی/ توئیت‌های فارسی

(Elasticsearch, Kafka, Cassandra, Spark, Redis)

مهلت تحویل : ۳ مرداد ماه ۱۴۰۰

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 :

۱. صف‌های توزیع شده و نقش محوری آن‌ها در سامانه‌های نوین اطلاعاتی.
۲. الاستیک سرچ و قدرت و کارآیی فوقالعاده آن در مدیریت داده‌های متنی و json
۳. کاساندرا به عنوان یک دیتابیس سطرگسترده مقیاس‌پذیر سهل الوصول و کارآمد
۴. اسپارک و سهولت پیاده‌سازی الگوریتم‌های پیچیده یادگیری ماشین بر روی حجم عظیم داده به کمک آن.

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

سعی شده است تمرکز اصلی پروژه، کار با ابزار و کتابخانه‌های ذکر شده باشد و خود کارهای پردازشی و کدهای مورد نیاز، حجم کمی را به خود اختصاص دهد.

چشم‌انداز کلی سامانه

در این پروژه قرار است داده‌های حدود هزارکانال اطلاع رسانی از پیام‌رسان‌های داخلی و یا توئیت‌های فارسی را به صورت لحظه‌ای بررسی کنیم و ضمن استخراج و ذخیره اطلاعات مفید از آن‌ها، بتوانیم برآوردی از زمان پست‌های بعدی آن‌ها و یا تعداد اشتراک گذاری آن‌ها داشته باشیم .

منابع اصلی ورود داده در این پروژه از قرار زیر هستند که می‌توانید یکی از آن‌ها را به دلخواه انتخاب نمایید :

۱. پیام‌رسان‌های داخلی مانند سروش، آی گپ و بله خواهند بود.
کدهای خزش برای پیام رسان‌ها توسط خود اعضای تیم باید نوشته شود.

۲. توئیتر و داده‌های فارسی روزانه آن .

۳. توئیت‌ها و پیا مهای سایت‌های فارسی بورس ایران مانند سهامیاب و ره‌آورد۳۶۵

هدف عملیاتی این پروژه، بررسی امکان خزش و تحلیل داده‌های پیام‌رسان‌های داخلی و یا توئیت‌های فارسی، مانیتورینگ و یافتن داده‌های آماری مرتبط با هرکانال(در پیام‌رسان‌ها) و هشتگ (برای توئیت‌ها) و انجام پردازش‌های مختلف بر اساس داده‌های آن‌ها به صورت بلادرنگ و نمایش آن‌ها به کاربر از طریق داشبوردهای اطلاعاتی خواهد بود.

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

- داده‌ها، به کمک و بهوک یا API های هر پیام‌رسان یا توئیتر و سایت‌های فارسی بورس، دریافت و وارد کانال اولیه در کافکا می‌شوند. (هماهنگی کل پروژه و گام‌های مختلف از طریق کافکا انجام می‌شود که در دنیای واقعی هم همین نقش بر عهده این نرم‌افزار است)
- در گام اول(PreProcess)، پیش‌پردازش‌های اولیه متنی بر روی داده‌ها انجام شده، کلمات کلیدی و هشتگ‌ها استخراج می‌شوند و به عنوان متادیتا، در کنار داده‌های دریافت شده قرار می‌گیرند. این داده‌ها وارد کانال دوم می‌شوند .

- در گام دوم(persistence)، داده‌های دریافتی در الاستیک سرچ ذخیره شده، بدون انجام پردازش خاصی، وارد کانال سوم می‌شوند .
- در گام سوم(ChannelHistory)، داده‌ها براساس نام خبرگزاری یا ارسال کننده محتوا/توئیت، کلمات کلیدی ، هشتگ‌ها، اشخاص یا کلمات خاص، در کاساندرا ذخیره می‌شوند. هدف از این مرحله، ایجاد مکانیزمی برای بازیابی سریع پست‌ها براساس نام کانال، کلمه کلیدی، هشتگ یا اشخاص/کلمات خاص است. سپس داده‌ها وارد کانال بعدی می‌شوند.
- در گام چهارم، می‌خواهیم بتوانیم برخی مدل‌های پیشبینی‌کننده را با اتصال اسپارک به کاساندرا تولید کرده، گروه بندی خودکار (هشتگ زنی خودکار) و پیشبینی زمان ارسال پست بعدی هر کانال را هم انجام دهیم بعد از ایجاد مدل پیشبینی هشتگ، این مدل به گام پیش‌پردازش اضافه خواهد شد که کیفیت برچسب‌زنی و استخراج کلمات کلیدی پست‌ها، ارتقا یابد.
- در گام پنجم(Statistics)، اطلاعات آماری مورد نیاز مانند تعداد اخبار در یک حوزه خاص، خبرگزاری خاص، هشتگ خاص و مانند آن، به روز رسانی می‌شود. این اطلاعات در ردیس ذخیره می‌شود. (این بخش دارای امتیاز اضافی خواهد بود)

همزمان با دریافت داده‌ها، باید بتوان :

- انواع جستجوهای متنی را روی محتوای لحظه‌ای کانال‌ها درون الاستیک سرچ انجام داد .
- آمار لحظه‌ای داده‌ها توسط یک وب اپلیکیشن و با خواندن داده‌ها از ردیس، به کاربر نمایش داده شود.

در ادامه، هر یک از پنج گام پردازشی فوق و نیز الزامات کلی پروژه به تفصیل بیان خواهند شد .

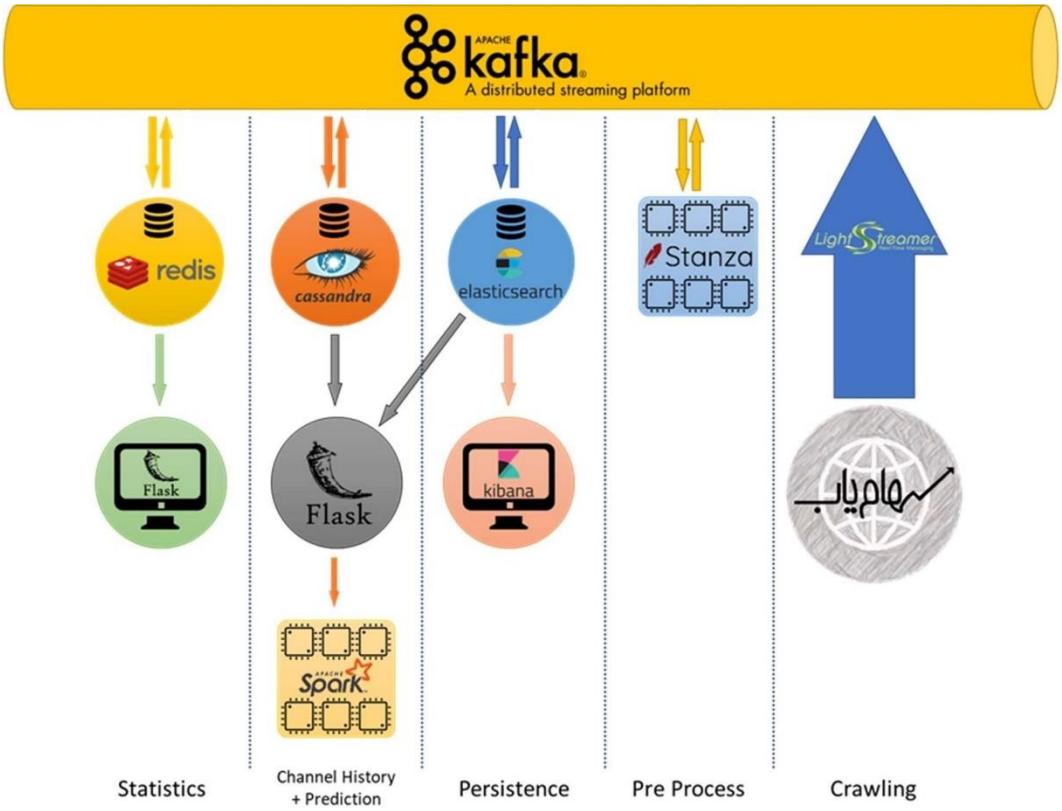
پیش‌نیازها و توضیحاتی در مورد ابزار و کتابخانه‌ها

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) 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 تسک‌ها و کارها را انجام دهید.

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



روال پیشنهادی تقسیم کار

در این پروژه به مهارت‌ها و کارهای زیر نیاز است :

- خواندن اطلاعات از پیام‌رسان و ارسال لحظه‌ای آن‌ها به کافکا (و ساخت کانال‌های مختلف کافکا).
- پردازش اولیه متن و ذخیره اطلاعات استخراج شده در الاستیک سرچ و نمایش آن‌ها در یک داشبورد درون کیبانا، نیز ذخیره اطلاعات آماری درون ردیس و نمایش آن‌ها به کمک یک داشبورد وب که با فلسک می‌تواند پیاده‌سازی شود.
- ذخیره اطلاعات تاریخچه‌ای درون کاساندرا و ساخت یک مدل پیشبینی کننده زمان پست بعدی هر کانال و دسته بندی هر متن (هشتگ زنی خودکار) با اتصال اسپارک به کاساندرا .

می‌توانید برای تقسیم کار بین اعضای تیم از بخش‌بندی فوق استفاده کنید.

نحوه تحویل کار

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 آنلاین، 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 .

استفاده از یک سرور (فیزیکی یا vps) و تحویل آنلاین پروژه، نمره امتیازی خواهد داشت .

گام اول : دریافت اطلاعات و Preprocess

برای دریافت اطلاعات از پیام‌رسان‌ها، از خزشگرهایی که توسط یکی از اعضای تیم نوشته خواهد شد استفاده کنید.

این اطلاعات به صورت مداوم از طریق برنامه‌ای که به صورت مداوم در حال اجراست و یا از طریق فراخوانی مداوم API ،به صورت json وارد کانال PreProcess کافکا خواهد شد.

انتظار می‌رود با نوشتن یک بات و عضو کردن آن در کانال‌های مختلف، به محض ارسال یک پست جدید در یک کانال، اطلاعات آن به سامانه پردازشی منتقل شود. کافی است عبارت «ساخت بات برای سروش/بله/آی‌گپ» را سرچ کنید تا بتوانید باتی برای خزش اطلاعات هر کانال طراحی کنید. بعد از ساخت این بات، لیستی از کانال‌ها تهیه کرده و این بات را به عضویت آن‌ها درآورید.

برای توئیت‌های داخلی می‌توانید از روش‌های مختلفی مانند فراخوانی API ،Crawling و مانند آن استفاده کنید.

داده‌های توئیتر نیز با فراخوانی API های استریمینگ آن، به راحتی قابل دریافت است.

با دریافت اطلاعات هر پست / توئیت از طریق کانال PreProcess، فرآیند پردازش ما شروع می‌شود. ابتدا تایم استمپ زمان دریافت و یک UUID به عنوان شناسه منحصربفرد هر پست/توئیت به آن اضافه کنید. سپس هشتگ‌ها یا کلمات کلیدی آن را استخراج کرده و به عنوان متادیتا به اطلاعات دریافت شده، اضافه کنید. اگر متن، حاوی لینک است، لینک‌های آن استخراج شده و درون یک آرایه جداگانه قرار گیرد. (متن اصلی را هیچ گاه تغییر نمی‌دهیم فقط اطلاعات مورد نیاز را استخراج و به صورت جداگانه ذخیره کنید)

برای استخراج کلمات کلیدی/ هشتگ، می‌توانید ایست واژ ه‌ها و افعال را حذف کنید، سپس کلماتی که tf/idf بالاتری دارند را به عنوان کلمه کلیدی در نظر بگیرید. توضیح اینکه هر پست می‌تواند یک یا چند هشتگ داشته باشد که آن‌ها را درون فیلد Hashtags ذخیره خواهید کرد. اما چه این هشتگ‌ها را داشته باشد چه نداشته باشد، شما باید خودتان کلمات کلیدی را استخراج و درون فیلد Keywords ذخیره کنید.

در این مرحله اگر متن دریافت شده حاوی کلمات زیر بود، این کلمات حتما به عنوان کلمات کلیدی باید درون آرایه Keywords قرار گیرند:

- **بورس – اقتصاد – تحریم – دولت – حسن روحانی**
- **انتخابات – دلار – طلا – کرونا**
- **کوید۱۹ (به هر شکل که نوشته شود) – تورم – دانشگاه**

در انتهای این مرحله یک json کامل از داده دریافت شده (داده‌های اصلی + متادیتای ایجاد شده) تولید می‌شود که آماده ذخیره‌سازی و پردازش‌های بعدی است. این متن وارد کانال persistence در کافکا خواهد شد.

گام دوم - persistence

در این مرحله، داده‌های دریافت شده مرحله قبل در الاستیک سرچ ذخیره می‌شوند.

دقت کنید که برای متون فارسی از Persian Analyzer[1] استفاده کنید. اگر بتوانید لیست ایست واژه‌ها و حتی Tokenizer را هم به صورت سفارشی (مثلاً استفاده از کتابخانه هضم در پردازش متون فارسی)، به الاستیک سرچ بدهید، امتیاز بیش‌تری خواهید گرفت.

داشبوردی در کیبانا طراحی کنید که موارد زیر را بتوان در آن مشاهده کرد :

- ابر کلمات یک کانال یا خبرگزاری خاص در یک بازه زمانی
- متن ده پست اخیری که دریافت شده است.
- تعداد پست‌های ارسال شده به ازای چند تا از کلمات کلیدی خاص که در مرحله قبل مشخص شده است در یک بازه زمانی.
- ده هشتگ بیش‌تر استفاده شده در پست‌های یک کانال خاص (یا تمام کانال‌ها) در یک بازه زمانی با تعداد تکرار هر هشتگ (یک نمودار ستونی) مثلا هشتگ‌های بیش‌تر استفاده شده در یک روز اخیر.
- یک نمودار به انتخاب خودتان.

ضمنا در گزارش قید کنید که اگر به دنبال تمام پست‌های حاوی یک کلمه خاص از یک خبرگزاری یا کانال خاص در یک بازه زمانی مشخص هستیم، چه دستوری باید بنویسیم. (و یا یک هشتگ خاص یا یک کاربر خاص در توئیت‌ها)

اگر تعداد پست‌ها/توئیت‌های ارسالی به ازای یک کلمه خاص را به ازای هر کانال / یا یک هشتگ خاص در توئیت‌ها در یک بازه زمانی بخواهیم، چه دستوری باید استفاده کنیم. (این کلمه، می‌تواند هر کلمه‌ای در متن باشد و ممکن است جزء کلمات کلیدی هم نباشد)

[1] https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-lang-analyzer.html

سوم - Channel/Hashtag History

در این مرحله، می‌خواهیم به کمک کاساندرا و مکانیزم ذخیره‌سازی سطرگسترده آن، تاریخچه زمانی هر کانال و هر هشتگ/کلمه کلیدی را ذخیره کنیم.

اگر کاربر نیاز داشت پست‌های اخیر یک کانال یا یک هشتگ را ببیند، کافی است داده‌ها از این دو جدول کاساندرا، خوانده شده و به کاربر نمایش داده شود. با توجه به اینکه کاساندرا، هنگام ذخیره سازی، داده‌ها را به صورت مرتب (طبق تنظیماتی که در تعریف جدول آورده ایم)، ذخیره می‌کند و از طرفی، عملیات جوین و اتصال هم نداریم، سرعت بسیار بالایی در واکشی اطلاعات دارد.

- دقت کنید که در کاساندرا، تکرار داده‌ها یک اصل کاملا پذیرفته شده است و به دنبال نرمال‌سازی نباشید.

- حداقل یک جدول برای کل پست‌ها (که بهتراست کلید هر سطر رو ز/ساعت دریافت هرپست باشد)، یک جدول برای هر کانال، یک جدول برای هر هشتگ/کلمه کلیدی نیاز خواهید داشت.

- کافی است فقط شناسه هر پست ذخیره شود. بعد از بازیابی اطلاعات مورد نیاز کاربر از کاساندرا، هنگام ارسال اطلاعات به کاربر، با دادن شناسه پست به الاستیک سرچ، اطلاعات کامل آن را می‌توانید بازیابی کرده و به کاربر نشان دهید.(نوع جستجوی ids در الستیک برای همین منظور ایجاد شده است) یعنی دراین پروژه از کاساندرا بیشتر به عنوان یک اندیس سفارشی شده روی داده‌ها استفاده خواهیم کرد.

نکته : تمام این اطلاعات را الاستیک 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.

در پایان این مرحله، داده‌ها وارد کانال Statistics می‌شود. (گام پنجم که اختیاری است)

انواع دستوراتی که برای بازیابی پست‌ها در یک ساعت اخیر، پست‌های یک کانال در ۲۴ ساعت اخیر، پست‌های مرتبط با یک هشتگ در بازه زمانی باید اجرا کنیم را هم در گزارش ذکر کنید.

آیا می‌توانیم اطلاعات آماری هر کانال ، هر هشتگ یا کل پست‌ها را در یک بازه زمانی به کمک کاساندرا به دست آوریم ؟

مثلا تمام پست‌های روزانه یک کانال در یک هفته گذشته ؟ پست‌های ذخیره شده در ماه گذشته ؟

چهارم - ساخت یک مدل پیشبینی کننده با اسپارک

با اتصال اسپارک به کاساندرا و استفاده از بخش MLIB آن، دو مدل برای پیشبینی موارد زیر بسازید :

- پیشبینی زمان ارسال پست بعدی یک کانال با دادن یک زمان خاص در یک روز خاص از هفته. مثلا ساعت هشت روز جمعه را به مدل می‌دهیم و انتظار داریم زمان ارسال پست بعدی به دقیقه را به ما بدهد.

- پیشبینی هشتگ‌های یک پست/توئیت. به ازای هر پست/توئیت و کلمات موجود در آن، کلمات کلیدی آن توسط این مدل، پیشبینی شود. البته برای این منظور، ابتدا باید پست‌ها/توئیت‌های زیادی که خود حاوی هشتگ باشند را دریافت کنید و سپس مدل را طوری آمو زش دهید که با دیدن یک مجموعه کلمات (یعنی هر پست) ، یک یا چند کلمه پیشنهادی برای آن، به عنوان نتیجه برگرداند. (این مدل اختیاری و دارای نمره اضافی می‌باشد)

می‌توانید از هر روش مکاشفه‌ای که بهبود دقت مدل‌ها کمک کند، استفاده کنید.

پنجم – Statistics (بخش امتیازی)

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

در این مرحله، اطلاعات آماری سامانه را به روز رسانی می‌کنیم.

به ازای هرکانال و هر هشتگ 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، پست‌های قدیمی باید از آمار فعلی کسر شوند، بنابراین درطراحی کلیدهای ردیس دقت به خرج دهید. به ازای هر پست یا مطلب جدیدی که دریافت می‌کنید، چندین کلید را در ردیس باید به روز رسانی 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 پست‌ها /توئیت‌های ارسال شده یک کانال خاص در شش ساعت گذشته.
- تعداد کل پست‌ها/توئیت‌های دریافت شده در یک بازه زمانی مثلا روز گذشته.
- تعداد هشتگ‌های دریافت شده در یک ساعت گذشته. (به صورت منحصر بفرد)
- آخرین هشتگ‌های دریافت شده. (یک لیست هزارتایی که با ورود داده‌های جدید، قدیمی‌ها حذف می‌شوند)
- آخرین پست‌ها/توئیت‌های دریافت شده (یک لیست صدتایی مشابه فوق)

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.

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.

ردیس در این پروژه برای به روز رسانی آمار لحظه‌ای استفاده می‌شود که برای این آمارها، نیاز به کوئری زدن به دیتابیس‌های مختلف نداشته باشیم.

نکات تحویل

- مهلت ارسال تا ۳ مرداد ماه خواهد بود .
- در این تمرین فقط مجاز به استفاده از زبان برنامه نویسی Python خواهید بود.
- انجام این تمرین به صورت تیمی می‌باشد و اعضای گروه می‌بایست در صورت سوال به یکدیگر کمک کنند.
- به صورت آنلاین و از طریق اسکایپ یا شاتل این پروژه تحویل گرفته خواهد شد که زمان آن بعدظهر همان روز از ساعت ۵ تا ۸ خواهد بود.
- گزارش شما در فرآیند تصحیح از اهمیت ویژه‌ای برخوردار است، لطفا تمامی مواردی که در شرح تمرین از شما خواسته شده را در گزارش ذکر نمائید.
- لطفا گزارش ، فایل کدها و سایر ضمائم مورد نیاز را با فرمت زیر ارسال نمائید. (هر نفر بخش مرتبط با خود / مدیر تیم ، کل گزارش)

Project_[Lastname]_[StudentNumber].zip