Elasticsearch, Kafka, Cassandra, Spark, Redis, Clickhouse, Superset

:

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 بیگ‌دیتا است. انتظار می‌رود پس از انجام این پروژه دیدی تجربی و شهودی نسبت به مفاهیم زیر پیدا کنید :

1. صف‌های توزیع شده و نقش محوری آن‌ها در سامانه‌های نوین اطلاعاتی.
2. الاستیک‌سرچ و قدرت و کارآیی فوق‌العاده آن در مدیریت داده‌های متنی و جی‌سان
3. کاساندرا به عنوان یک دیتابیس سطرگسترده مقیاس‌پذیر سهل‌الوصول و کارآمد
4. اسپارک و سهولت پیاده‌سازی الگوریتم‌های پیچیده یادگیری ماشین بر روی حجم عظیم داده به کمک آن .
5. سوپرست به عنوان یک ابزار دم‌دستی و کاربردی برای بصری سازی نتایج پردازش و ساخت داشبوردهای تحلیلی
6. دیتابیس‌های تحلیلی و نقش آن‌ها در تصمیمات مدیریتی سازمانی

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 ده‌هزارکانال اطلاع‌رسانی 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 اشتراک‌گذاری آن‌ها داشته باشیم .

با توجه به حجم کار این پروژه، می‌توانید تیم‌های حداکثر چهار نفره تشکیل دهید که هر تیم یک مدیر یا هماهنگ کننده خواهد داشت. در صورتی که تعداد اعضای تیم شما حداکثر دونفره باشد، با هماهنگی با دستیاران آموزشی می‌توانید از انجام بخشی از کار، صرف نظر کنید.

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

1. پیام‌رسان‌های داخلی مانند سروش، آی‌گپ و بله خواهند بود که هر تیم، با یکی از آن‌ها کار خواهد کرد. کدهای خزش برای پیام‌رسان‌ها توسط خود اعضای تیم باید نوشته شود.
2. توئیتر و داده‌های فارسی روزانه آن .
3. توئیت‌ها و پیام‌های 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 وب‌هوک 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 محتوی/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 (*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. سپس داده‌ها وارد کانال پنجم می‌شوند.
- در گام پنجم (*Analytics*)، داده‌های دریافت شده به غیر از خود متن دریافت شده، برای مقاصد تحلیلی وارد کلیک‌هوس می‌شوند و چرخه پردازش داده به اتمام می‌رسد.

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.
- انواع گزارش‌ها پیچیده بااتصال سوپرست به کلیک‌هوس، در لحظه قابل تولید و نمایش باشد.

علاوه بر اینها، می‌خواهیم بتوانیم برخی مدل‌های پیش‌بینی کننده 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. (**این بخش دارای امتیاز اضافی خواهد بود**). بعد از ایجاد مدل پیش‌بینی هشتگ، این مدل به گام پیش‌پردازش اضافه خواهد شد که کیفیت برچسب‌زنی و استخراج کلمات کلیدی پست‌ها، ارتقا یابد.

هر چند تأکید اصلی پروژه بر استفاده از پیام‌رسان‌های داخلی مانند سروش، بله ، آی‌گپ و مانند آن‌ها است اما برای شروع کار می‌توانید از داده‌های توئیتر استفاده کنید و پس از ساختن سامانه اصلی، منبع دریافت داده آنرا تغییر دهید.

برای استفاده از داده‌های توئیتر، می‌توانید از این آموزش (https://bit.ly/2YOiN5U) استفاده کنید و کلیدهای زیر را برای اتصال به توئیتر به کار برید :

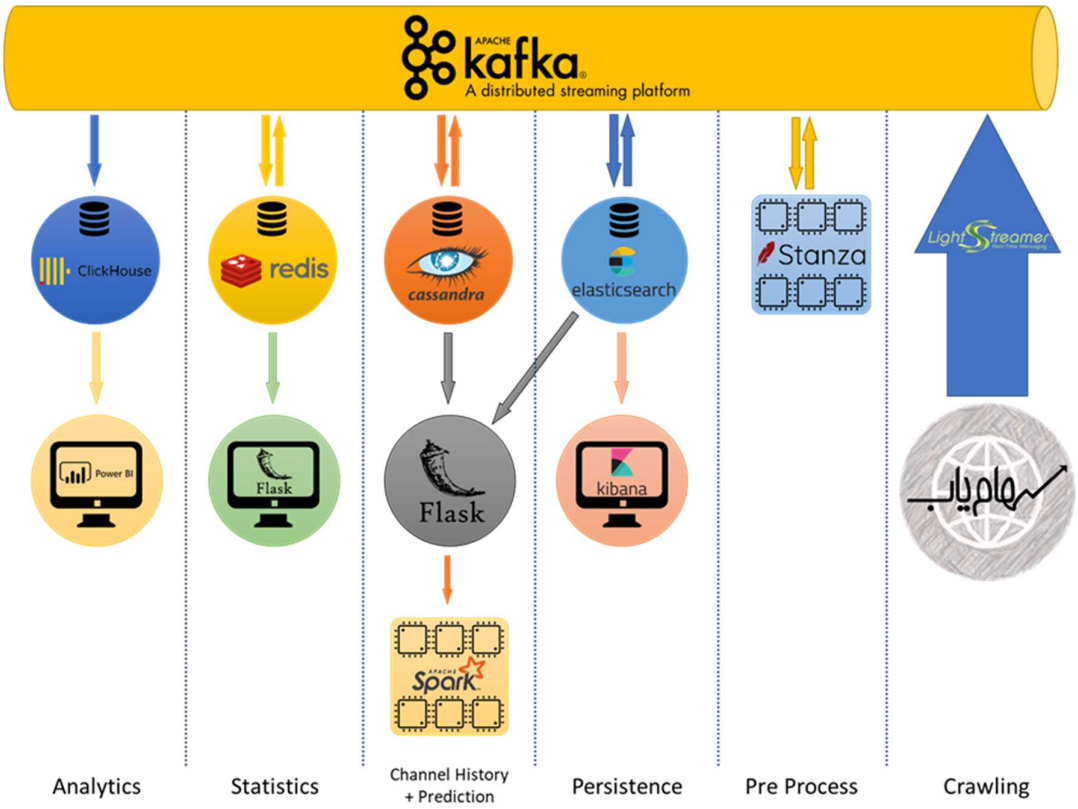
```
consumer_key = '2QX1YKQKheOsezZgotXZoiBXc'
consumer_secret = 'XWDOXGljhAP03SU1xweQS6PmoegvPBnHHdFwAadG6CnPTnWHjK'
access_token = '15257539-ERDMc7Ezn7t0tLmfBRRruyYpGmIsN43hsGSHdQS64'
access_secret = '1DH7FHDcqgHX3YxW2ZcvU91dkaZcogISXUevCw1PxScoQ'
```

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. **پیشنهاد ما استفاده از کافکا است اما میتوانید از RabbitMQ یا NSQ هم استفاده کنید**. تعداد اعضای هر تیم، بهتر است دو تا سه نفر باشد اما گروه‌های چهار نفره هم مجاز خواهد بود. بهتر است برای هماهنگی بیشتر، یک نفر را به عنوان مدیر تیم انتخاب کرده، هماهنگی و توزیع تسک‌ها و کارها را از طریق گیت‌لب/گیت‌هاب و از طریق مکانیزم برنچینگ و ایجاد ایشو انجام دهید.

شکل زیر شماتیک معماری این سیستم را که توسط یکی از تیم‌های سالهای گذشته این درس طراحی شده است 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 پیام رسان 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 پست‌بعدی هر کانال و دسته بندی هر متن (هشتگ زنی خودکار) با اتصال اسپارک به کاساندرا.
- ذخیره اطلاعات تحلیلی درون دیتابیس کلیک‌هوس و اتصال آن به سوپرست و ساخت چندین داشبورد تحلیلی درون سوپرست

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

نحوه تحویل کار

گزارش نهایی پروژه توسط مدیر تیم در ایلرن به همراه آدرس ریپوزیتوری گیت پروژه(در صورت وجود)، آپلود خواهد شد. **هر فرد از اعضای تیم، گزارش آماده شده برای بخش خودش را در سامانه آپلود خواهد کرد** تا در صورت کم‌کاری 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 به دستیاران آموزشی 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 جی‌سان 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 استریمنگ 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 آنرا استخراج کرده و به عنوان متادیتا به اطلاعات دریافت شده، اضافه کنید. اگر متن، حاوی لینک است، لینک‌های آن استخراج شده و درون یک ارایه 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 ایست‌واژه‌ها 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 حتماً 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 ذخیره سازی 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 ایست‌واژه‌ها 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) مثلاً هشتگ‌های بیشتر استفاده شده در یک روز اخیر .
- یک نمودار به انتخاب خودتان.

ضمناً 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. (ویا 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