بسم الله الرحمن الرحیم

موضوع پروژه :

**ساختن سیستم های پیچیده اطلاعاتی و نظارت و مانیتورینگ مداوم**

استاد مربوطه : مهندس سید مجتبی بنائی

گردآورندگان : جواد آسوده ، جوادسالاری ، محمد کیوانی ، سید احمد ملایی

توضیحات کلی پروژه :

ابتدا اطلاعات صفحه همایش ها به وسیله کتابخانه request یا urllib گرفته میشود و توسط این کتابخانه ها ما میتوانیم محتوای صفحه html را واکشی کنیم . طرز کار این کتابخانه در این پروژه بدین صورت است که ابتدا لینک لینک صفحه کنفرانس را به ورودی توابع مربوطه دادیم و پس از دادن خروجی html به عنوان ورودی به خزنده اول میدهیم که اولین لینک همایش ها را پیدا نموده و آن را به صف کافکا میفرستیم . این کار مکررا انجام شده تا تمام لینکهای همایش کرول شده و به ترتیب وارد صف کافکا شوند . از طرف دیگر همیشه مصرف کننده ای وجود دارد که به این صف کافکا که در پورت خاصی در حال گوش دادن است و اطلاعات موجود در آن را برمیدارد ، برای مثال اولین لینکی که وارد صف شده است را گرفته و دوباره آن را با استفاده از خزنده دوم کرول کرده و لینک های مقالات موجود در آن همایش خاص را بدست می آورد . این مرحله نیز همانند مرحله قبل است زمانیکه اولین مقاله آن همایش را استخراج میکند آن را به topic دیگری در همان صف ارسال میکند و به همین روش تمام لینک های مقالات درون این topic قرار میگیرند . مصرف کننده دیگری وجود دارد که به این topic همیشه گوش میدهد، در این هنگام ابتدا لینک مقاله مورد نظر بررسی میشود که در دیتابیس riak یا redis وجود داشته است یا نه ، اگر لینک مقاله مورد نظر در دیتابیس موجود بود مقاله ثبت نشده و مقاله بعدی مورد بررسی قرار میگیرد ، ولی اگر لینک مقاله در دیتابیس (riak یا redis) موجود نبود لینک این مقاله در دیتابیس ثبت شده و دوباره توسط کتابخانه urllib تمام اطلاعات این مقاله کرول میشود و همه این اطلاعات درون الاستیک سرچ ذخیر میشوند . در هر یک از مراحل بالا ممکن است سیستم دچار خطا شود، در این حالت ما تمام خطاها را لاگ نموده و آنها را به ترتیب در یک topic دیگری درون صف کافکا قرار میدهیم ، تا این اطلاعات به مرحله ی تحلیل رفته و آماده مانیتورینگ شوند . زمانیکه اطلاعات درون الاستیک سرچ ذخیره شد ، بعد از آن ما این اطلاعات را به صورت یک topic دیگری درون صف کافکا قرار میدهیم ، در این مرحله نیز مصرف کننده ای مربوط به بخش تحلیل وجود دارد که این اطلاعاتی که درون صف کافکا قرار دارند را درون دیتابیس کاساندرا ذخیره میکند . در زمانی که کاربر قصد دارد مقاله مورد نظر خود را جستجو کند ، ما دو روش برای ارائه این اطلاعات پیش ور داریم . اگر کاربر نام مقاله ای خاص و یا قسمتی از متن مقاله را جستجو کند ما میتوانیم به سرعت اطلاعات را بوسیله الاستیک سرچ جستجو کرده و به کاربر نمایش دهیم . و یا اگر کاربر بخواهد بر اساس فیلد ها و مقادیر خاص جستجوی خود را انجام دهد (برای مثال : پیدا نمودن تمام مقالاتی که در سال 1394 منتشر شده و دارای برچسب ریاضی هستند) ما با استفاده از دیتابیس کاساندرا اطلاعات مورد نیاز کاربر را بسرعت برمیگردانیم ، در این سیستم صفحه ای بنام گزارش گیری وجود دارد که کاربر در این صفحه میتواند لیست لاگ هایی که در مرحله قبل وارد صف شده بودند و به بخش تحلیل ارسال شده اند را ببیند و از مشکلات سیستم آگاه شود ، و همینطور میتوانیم به کاربر نشان دهیم که چه مقالاتی بدلیل تکراری بودن زیاد از اهمیت بالاتری برخوردارند و چه مقالاتی اهمیت کمتری دارند . برای نمایش لاگ ها در صفحه به محض وارد شدن کاربر به صفحه نمایش لاگ ها ارتباط socket برقرار میشود و تابع open سمت کدهای backend فراخوانی شده و در آن هنگام ابتدا مصرف کننده ما در صورت وجود پیام در topic لاگ پیام را به سمت کلاینت تا موقع خالی شدن صف و بسته نشدن ارتباط socket ارسال میکند .