توسعه یک برنامه جریان داده با استفاده از Spark Structured Stream

سلام
برای آشنایی بیشتر با مدل کاری پروژه‌هایی که در شرکت اجرا می‌شوند، تسکی برای شما تعریف شده که بتونید با این فضا بیشتر آشنا بشی.

نحوه تحویل پروژه

کلیه ی کدها و فایلهای تهیه شده برای انجام این پروژه می بایست در یک control source گیت قرار گرفته شود. پس از انجام هر بخش از پروژه یک کامیت در پروژه داشته باشید.

هدف پروژه

هدف از این پروژه پیاده سازی یک آنومالی دیتکشن ساده با استفاده از Spark است.

مرحله 1

در این مرحله داده ها با استفاده از یک Bash Script از فایلی که در اختیار شما قرار داده شده خوانده می‌شود و به کافکا ارسال می‌شود.
اپلیکیشن استریمینگ مورد نظر می‌بایست دیتا را از رویtopic موردنظر کافکا خوانده و محاسبات لازم را انجام دهد. در صورت بروز آنومالی در دو سطح Warning و Error پیام مناسب در کنسول چاپ می شود.
برای تشخیص آنومالی، میانگین و انحراف معیار داده ها در یک بازه معقول ذخیره شده و در صورتی که در پنجره کنونی میانگین متریک مورد نظر به اندازه یک انحراف معیار از میانگین کلی داده هافاصله داشته باشد به عنوان سطح Warning و چنان چه بیش از دو انحراف معیار فاصله داشته باشد به عنوان سطح Error در نظر گرفته می‌شود.

خروجی های این مرحله

- فایل bash برای خواندن دیتا و ارسال به کافکا
- کد اپلیکیشن نوشته شده به همراه jar فایل
- ارائه مستندی کوتاه از نحوه اجرا برنامه

مرحله 2

در این مرحله میخواهیم سرویس های خود را به صورت کانتینرهای داکر اجرا کنیم. برای این کار می توانید از یک داکر ایمیج برای اجرای standalone اسپارک استفاده کنید. مواردی که در این مرحله اهمیت دارد نحوه تنظیم کانفیگها و اجرای پروژه انجام شده در محیط کانتینری است

خروجی‌های این مرحله

- کد compose-docker برای اجرا شدن سرویس آنومالی دیتکشن. با اجرای دستور compose-docker up- it می بایست اسپارک به صورت standalone اجرا شده و برنامه استریمینگ مورد نظر شروع به اجرا کند. خروجی های برنامه نوشته شد در این حالت می‌بایست در کنسول/فایل قابل مشاهده باشد.
- ارائه مستند کوتاه در مورد روند انجام پروژه.