UnB/FCTE – Engenharia de Software
PSPD - Programação para Sistemas Paralelos e Distribuídos Prof. Fernando W. Cruz

Atividade extraclasse sobre Spark/Kafka (com API Gráfica)

A) Descrição e objetivos

O objetivo deste experimento é que o aluno compreenda como configurar uma aplicação Spark Streamming (https://spark.apache.org/docs/latest/streaming-programming-guide.html) para processar dados obtidos em tempo real e gerar resultados em *dashboards* gerenciais.

B) Roteiro do laboratório

Para alcançar os objetivos propostos, os alunos devem criar uma nova versão do 4o. Notebook Google Colab disponível no Moodle da disciplina (que está em https://colab.research.google.com/drive/1BHFbFP7Bs38APEhYOsOxqrEYFNbXbBSt?usp=drive_link), de modo a substituir o processamento, as entradas e as saídas, da seguinte forma:

- Entrada: Substituir o esquema atual de entrada por um método de coleta de palavras a partir de alguma rede social, como o Discord, Twitter/X ou outra rede que suporte o consumo via canais kafka. Nesse caso, todo o processo de configuração e consumo das palavras da rede social, bem como os comandos de recuperação dos dados de entrada devem estar documentados no Notebook Google Colab (ou em documento à parte, em último caso). Caso o uso de rede social como entrada não seja possível, os alunos devem escrever texto justificando os problemas que tiveram e o que os motivaram a mudar de estratégia para viabilizar a entrega do *notebook*.

- Processamento: Substituir o contador de palavras do *notebook* original por algum processamento que envolva o consumo de API de uma das *engines* de IA (Inteligência Artificial) disponíveis, tais como OpenAI (https://openai.com/), Google Gemini (https://gemini.google.com/) e DeepSeek (https://www.deepseek.com/, se disponível). Embora o tipo de aplicação a ser entregue seja de livre escolha por cada grupo, seguem algumas dicas:
  a. Priorizar aplicações que não envolvam custos de utilização dos *engines* de IA.
  b. Incluir documentação sobre a API do *engine* de IA escolhido, com foco nos passos necessários para seu uso/consumo.
  c. Além da interação com o *engine* de IA, se a solução apresentada envolver algum modelo de rede neural diferenciado, este deve ser devidamente documentado. Ou seja, a entrega deve incluir explicação adicional sobre o modelo usado. Essa descrição pode ser feita no próprio *notebook* que será entregue, de forma sintética, porém clara.
  d. Um exemplo de aplicação compatível com os objetivos dessa atividade extraclasse é o desenho de um engine Spark Streaming que recebe, via canal kafka, perguntas/ percepções dos usuários (mensagens de uma rede social, em tempo real) e as envia imediatamente para o *engine* de IA (talvez algum tratamento da mensagem antes de envio seja necessário). As respostas, por sua vez, são consolidadas e apresentadas no *dashboard* de saída.

e. A proposta de processamento deve ser original, mas pode ser embasada em alguma solução já pronta. Nesse caso, é preciso indicar a referência usada para confecção da entrega solicitada neste trabalho extraclasse.

- Saída: Substituir o canal de saída atual via console por gráficos e *charts* usando ElasticSearch (https://www.elastic.co/pt/elasticsearch) e Kibana (https://www.elastic.co/pt/kibana) integrados a um canal kafka de saída de modo que, tudo o que for escrito no canal de saída seja visualizado em gráficos, no Kibana. Todos os passos de instalação e configuração necessários para viabilizar esse tipo de saída devem ser documentados no notebook Google Colab. Da mesma forma, se a utilização do Elastic/ Kibana não for possível como *dashboard*, os alunos devem escrever texto justificativo e adotar alguma outra alternativa gráfica compatível.
  Obs.: Além da instalação do elasticsearch e do kibana, sugere-se o uso do kafka Connect (https://docs.confluent.io/platform/current/connect/index.html), que é uma espécie de *plugin* que permite associar o canal kafka de saida do Spark (canalouput) ao elasticsearch/kibana. Depois dessa vinculação, o próximo passo é entrar no elasticsearch/kibana e definir os tipos de gráficos desejados para apresentação dos dados que chegarem pelo canal de saída (canaloutput).

C) Requisitos da entrega

O notebook a ser entregue deve ser feito como Google Colab (não utilizar Jupyter Notebook) e deve conter todos os comandos que garantam o alcance do objetivo proposto, sem a necessidade de uso de comandos externos ao próprio notebook (valem apenas os comandos do *notebook*). Ou seja, deve-se manter o mesmo roteiro do Google Colab usado como referência, alterando apenas as partes relacionadas aos tópicos citados na seção anterior. A sequência do *notebook* deve seguir (mais ou menos) a seguinte sequência:

1. Inicializações para o laboratório Google Colab
2. Configurações de mecanismos para garantir o acesso aos resultados do notebook (visualização dos resultados da contagem de palavras na forma gráfica)
3. Instalação/configuração do Spark - versão cluster, preferencialmente
4. Instalação/configurando do kafka e canais de entrada e saída
5. Instalação/configuração da API de consumo de rede social
6. Instalação/configuração de saída gráfica
7. Interação com *engine* de IA no Spark
8. Apresentação de resultados em um *dashboard* gráfico no Kibana/ElasticSearch

D) Questões de ordem

- A atividade pode ser feita por grupos de, no mínimo 3 e, no máximo 4 alunos, em turmas menores do que 50 alunos e de 4 a 5 alunos em turmas maiores. Nesse caso, basta que um dos alunos faça a postagem do material no Moodle (arquivo zipado)
- No material deve constar os nomes/matrículas dos membros do grupo, a fim de que sejam beneficiados com a avaliação.

- A entrega a ser feita é *(i)* o próprio notebook Google Colab devidamente documentado (incluindo identificação dos membros do grupo e da disciplina), e *(ii)* um vídeo de demonstração de preparação e uso do *notebook*, incluindo discussões e decisões tomadas pelo grupo até chegar a versão final entregue. Considerar 3 a 4 minutos por membro para produção do vídeo. Os arquivos (*notebook* e vídeo) devem ser compactados e enviados ao Moodle dentro dos limites de prazo estabelecidos pelo professor.
- Além do *notebook*, se houver algum passo adicional para uso e teste da instalação, esse deve constar no material de entrega, com as orientações devidas.
- Serão oferecidos pontos extras, se os alunos apresentarem funcionalidades adicionais (devidamente documentadas) além do que foi proposto.
- Incluir no final do *notebook* todas as referências utilizadas para realização do laboratório, incluindo vídeos e outros tipos de materiais na Internet, por exemplo.