Gestão de Grandes Conjuntos de Dados

Engenharia Informática – Universidade do Minho

Segundo Trabalho Prático – 2019/2020

O resultado do trabalho é o código fonte identificando claramente como atinge cada um dos objetivos e um relatório escrito. O relatório deve omitir considerações genéricas sobre as ferramentas utilizadas, focando a apresentação e justificação dos objetivos atingidos. A entrega do relatório é feita na área da Unidade Curricular no *e-Learning*, pelos grupos já constituídos. A data limite é 5 de junho de 2020.

1 Contexto

O trabalho prático consiste na concretização e avaliação experimental de tarefas de armazenamento e processamento de dados utilizando Spark (*batch* e *streaming*). Os dados a utilizar são:

- o *dataset* público do IMDB: `https://www.imdb.com/interfaces/`
- o gerador de *streams* baseado nos mesmos dados e que simula uma sequência de votos individuais de utilizadores, disponível no *eLearning*.

2 Objetivos

1. Uma componente de processamento de *streams* que produza os seguintes resultados:
   (a) *Log:* Armazene todos os votos individuais recebidos, etiquetados com a hora de chegada aproximada ao minuto, em lotes de 10 minutos. Cada lote deve ser guardado num ficheiro cujo nome identifica o período de tempo.
   (b) *Top3:* Apresente a cada minuto o top 3 dos títulos que obtiveram melhor classificação média nos últimos 10 minutos.
   (c) *Trending:* Apresente a cada 15 minutos os títulos em que o número de votos recolhido nesse período sejam superiores aos votos obtidos no período anterior, independentemente do valor dos votos.
2. Uma componente de processamento em *batch* que permita realizar as seguintes tarefas:
   (a) *Top10:* Calcule o top 10 dos atores que participaram em mais títulos diferentes.
   (b) *Friends:* Calcule o conjunto de colaboradores de cada ator (i.e., outros atores que participaram nos mesmos títulos).
   (c) *Ratings:* Atualize o ficheiro `title.ratings.tsv` tendo em conta o seu conteúdo anterior e os novos votos recebidos até ao momento.
3. Escolha a configuração e a implementação que, para o mesmo *hardware*, lhe permite receber e tratar o maior débito de eventos. Justifique com resultados experimentais.

3 Notas

- Tirando partido da *Google Cloud* deve usar mais do que um processo *worker* e armazenar todos os ficheiros no sistema HDFS.
- Descreva a configuração de *hardware* e *software* utilizada nas experiências que efetuar.
- Inclua todo o código-fonte e ficheiros de configuração necessários para executar os programas pedidos. Inclua no relatório instruções claras para a utilização destes programas.
- Justifique com argumentos objetivos as opções tomadas, tanto em termos de algoritmos como de parâmetros de configuração. Por exemplo, corra e compare medidas das alternativas sempre que achar necessário.
- Sugere-se as utilização das versões reduzidas dos dados (*mini* e *micro*) disponibilizadas no *eLearning* durante o desenvolvimento e testes, mas a resolução deve funcionar eficientemente com os dados completos. Os resultados experimentais relatados devem também considerar os dados completos.