Gestão de Grandes Conjuntos de Dados

Engenharia Informática – Universidade do Minho

Segundo Trabalho Prático – 2020/2021

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 4 de junho de 2021.

1 Contexto

O trabalho prático consiste na concretização e avaliação experimental de tarefas de armazenamento e processamento de dados utilizando Spark e Hive Metastore. Os dados a utilizar são o *dataset* público do IMDB: https://www.imdb.com/interfaces/

2 Objetivos

1. Carregue os ficheiros do *dataset* para os formatos mais eficientes e adequados às operações que vai efetuar.
2. Desenvolva operações com Spark RDDs que respondam às seguintes interrogações analíticas:
   (a) *Top Genres:* Género mais comum em cada década.
   (b) *Season Hits:* Título mais bem classificado em cada ano.
   (c) *Top 10:* Top 10 dos atores que participaram em mais títulos diferentes.
3. Desenvolva operações com Spark RDDs ou SQL que materializem num único ficheiro, com o esquema apropriado, os dados para "páginas de ator" contendo a seguinte informação:
   (a) *Base:* Nome, idade, número de títulos em que participou, intervalo de anos de atividade e classificação média dos títulos em que participa.
   (b) *Hits:* Top 10 dos títulos mais bem classificados em que participou.
   (c) *Generation:* Nomes do top 10 de atores da mesma geração (i.e., que nasceram na mesma década).
   (d) *Friends:* Conjunto de colaboradores de cada ator (i.e., outros atores que participaram nos mesmos títulos).

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 **Não inclua ficheiros de dados.**
- 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.