Gestão de Grandes Conjuntos de Dados

Engenharia Informática – Universidade do Minho

Primeiro Trabalho Prático – 2020/2021

O resultado do trabalho é o código fonte 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 8 de abril de 2021.**

1 Contexto

O trabalho prático consiste na concretização e avaliação experimental de tarefas de armazenamento e processamento de dados utilizando Hadoop HDFS, Avro+Parquet e MapReduce. Os dados a utilizar são o *dataset* público do IMDB: `https://www.imdb.com/interfaces/`

2 Objetivos

As tarefas de processamento de dados a realizar são as seguintes:

1. Carregue os dados dos ficheiros `title.basics.tsv.gz` e `title.ratings.tsv.gz` para um único ficheiro AvroParquet com um esquema apropriado.
2. Usando o ficheiro resultante da alinea anterior e considerando apenas filmes (*movie*), calcule para cada ano:
   - o número total de filmes;
   - o filme que recolheu mais votos;
   - os 10 melhores filmes segundo a classificação.

   Estes resultados devem ser armazenados num ficheiro AvroParquet com um esquema apropriado.
3. (Valorização) Para cada filme, recomende o outro do mesmo género que tenha a melhor classficação. Considere apenas o primeiro género de cada filme. Estes resultados devem ser apresentados em ficheiros de texto e deve evitar carregar em memória simultâneamente todos os filmes do mesmo género.[1]

3 Notas

- 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.

[1] Sugestão: `https://www.oreilly.com/library/view/data-algorithms/9781491906170/ch01.html`