Gestão de Grandes Conjuntos de Dados

Engenharia Informática – Universidade do Minho

Primeiro Trabalho Prático – 2019/2020

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 23 de abril de 2020.

1 Contexto

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

2 Objetivos

Deve ser entregue o código-fonte claramente identificado que realize cada uma das seguintes tarefas:

1. Carregue os dados do ficheiro `title.basics.tsv.gz` para uma tabela HBase.
2. Usando a tabela HBase da alinea anterior e os restantes ficheiros do *dataset*, compute os dados necessários para apresentar para cada ator uma página contendo:

   - nome, datas de nascimento e morte;
   - número total de filmes em que participou como ator;
   - títulos dos três filmes com melhor cotação em que participou.

   Estes dados devem ser armazenados numa tabela HBase.

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
- 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.