

## Bootcamp: Engenheiro de Dados Cloud

### Trabalho Prático

### Módulo 3: Soluções de Big Data e Data Lake

#### Objetivos de Ensino

Exercitar os seguintes conceitos trabalhados no Módulo:

1. Contexto de Big Data e das ferramentas de processamento massivo
2. Fundamentos do Spark.
3. Funcionamento interno do Spark.
4. Manipulação de dados com Spark.

#### Enunciado

Oi, pessoal!

Neste trabalho prático, vamos exercitar processos de leitura, limpeza e análise de dados sobre dois arquivos que são uma amostra dos dados disponíveis no IMDB.

O IMDB (Internet Movie Database) é um repositório online de informações sobre cinema, TV, música e jogos.

- *title\_basics.tsv*, que contém informações sobre cada filme, como nome do filme, duração e gênero;
- *title\_ratings.tsv*, que contém, para cada filme, sua nota média atribuída pelas pessoas que avaliaram o filme, bem como o número de pessoas que votaram.

Veja a descrição completa do conteúdo destes arquivos aqui: <https://www.imdb.com/interfaces/>

Baixe os arquivos em

<https://homepages.dcc.ufmg.br/~pcalais/XPE/engenharia-dados/big-data-spark/trabalho-pratico/>

Como estes dados são estruturados, utilize a API de Dataframes do Spark.

Para ler o arquivo de metadados do filme, use o seguinte comando no pyspark:

```
df_titles = spark.read.csv('title_basics.tsv', header=True, inferSchema=True, sep='\t')
```

Analogamente, para ler o arquivo de avaliações de filmes, use o seguinte comando:

```
df_ratings = spark.read.csv('title_ratings.tsv', header=True, inferSchema=True, sep='\t')
```

Dica: use a transformação join para criar um dataframe único, que contém os dados de ambos os arquivos.

Antes de executar o trabalho prático e responder às questões objetivas, assista às aulas dos Capítulos 1, 2 e 3 e consulte a apostila! Em particular, o capítulo com as instruções para instalação do Apache Spark! 😊

Os seguintes links têm um resumo dos principais comandos da API de Dataframes do Spark e podem ser úteis para você:

[https://s3.amazonaws.com/assets.datacamp.com/blog\\_assets/PySpark\\_SQL\\_Cheat\\_Sheet\\_Python.pdf](https://s3.amazonaws.com/assets.datacamp.com/blog_assets/PySpark_SQL_Cheat_Sheet_Python.pdf)  
<https://gist.github.com/AlessandroChecco/c930a8b868342fa34b23a1f282dc3e88>

Bom trabalho!

## Atividades

Os alunos deverão desempenhar as seguintes atividades:

1. Instalação do Apache Spark (local ou uso de solução em nuvem)
2. Leitura dos arquivos usando a API de Dataframes
3. Execução de algumas análises de dados usando a API de Dataframes do Apache Spark