# Bootcamp: Engenheiro(a) de Dados (Cloud)

## Trabalho Prático

## Módulo 2: Tecnologias de Big Data – Processamento de Dados Massivos

### 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

Durante a primeira metade do módulo, as aulas práticas utilizaram um conjunto de dados gratuito disponível no site do IMDB, com informações de filmes séries e outras produções audiovisuais. Neste trabalho prático, você deverá fazer o download das tabelas **title.basics** e **title.ratings** do site oficial (https://datasets.imdbws.com/) e realizar um processo de limpeza nos dados vistos em nas aulas, utilizando o Apache Spark. Será necessário alterar os tipos das colunas, tratar os valores nulos e realizar algumas análises com os dados.

### Atividades

Os alunos deverão desempenhar as seguintes atividades:

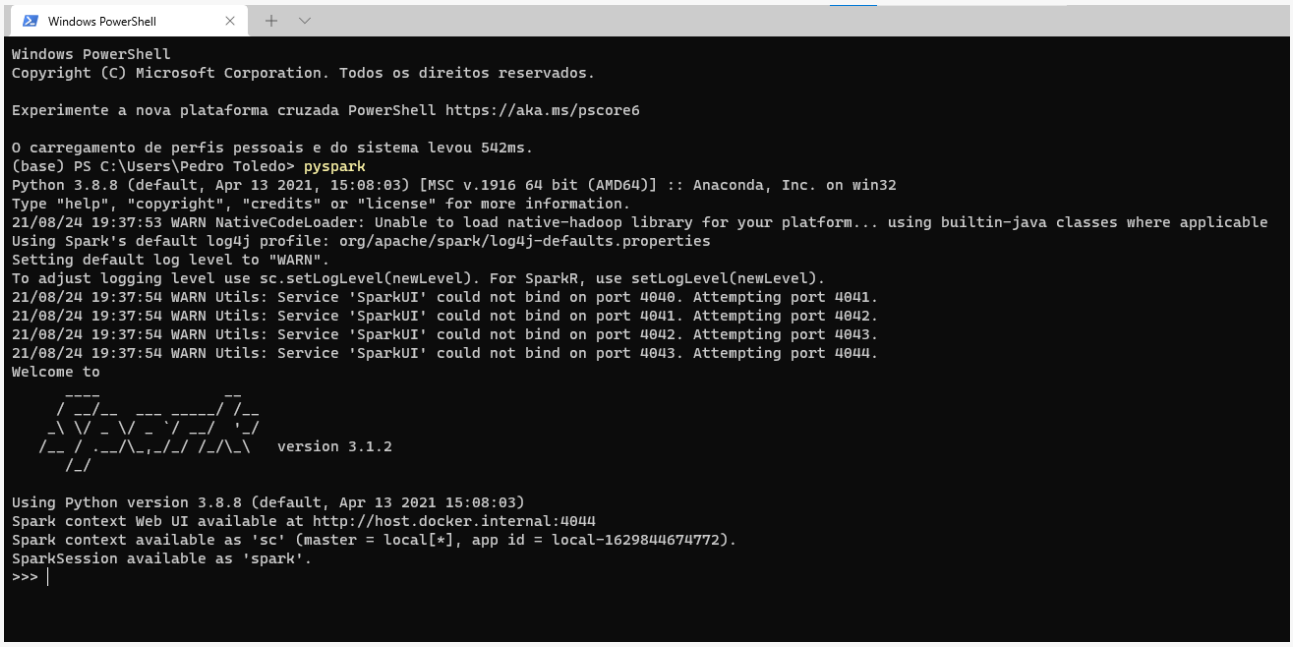
# 1 – Instalação do Apache Spark

Os alunos deverão assistir a aula 5 do capítulo 2: Download e instalação do Apache Spark e realizar os passos indicados para a instalação do programa. Ao final da instalação, será necessário abrir um terminal de linha de comando que possa executar ou o Python ou o Scala (Anaconda Prompt, Windows PowerShell etc.), e habilitar uma shell do Apache Spark, utilizando um dos seguintes comandos:

- Python: pyspark.
- Scala: spark-shell.

**Obs.:** É recomendado que os alunos utilizem a API do Python para manipular dados no Spark, uma vez que alguns dos comandos diferem ligeiramente entre as duas linguagens. Caso o aluno escolha utilizar Scala, será assumido que o aluno conhece a linguagem e tem capacidade de lidar com essas pequenas diferenças sozinho.

**Figura 1 – Criando a Shell do Pyspark**




Esse será o ambiente em que a prática será guiada. Se o aluno preferir, também poderá realizá-la em Jupyter Notebooks.

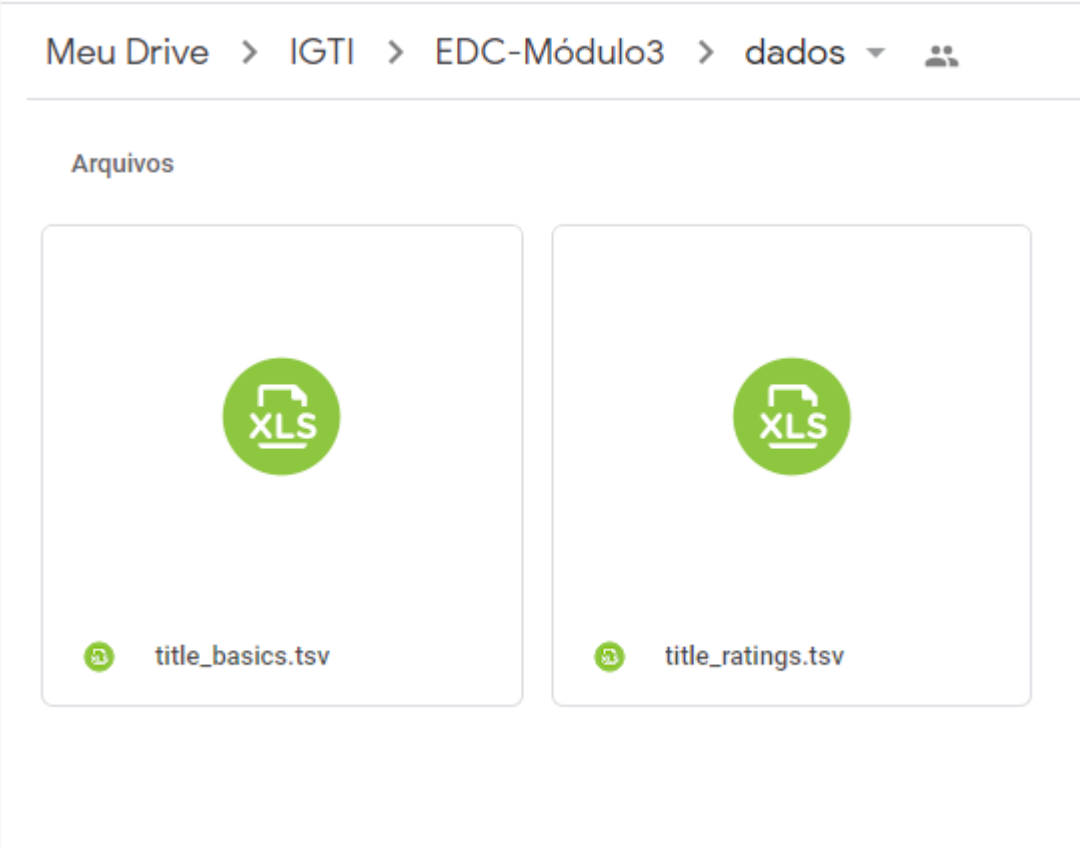
# 2 – Download dos Arquivos

Em seguida, será necessário fazer o download das tabelas no seguinte link:

https://drive.google.com/drive/u/1/folders/1DfuJmlOsXU8hgCRUui-89FQhhHh3L5kS

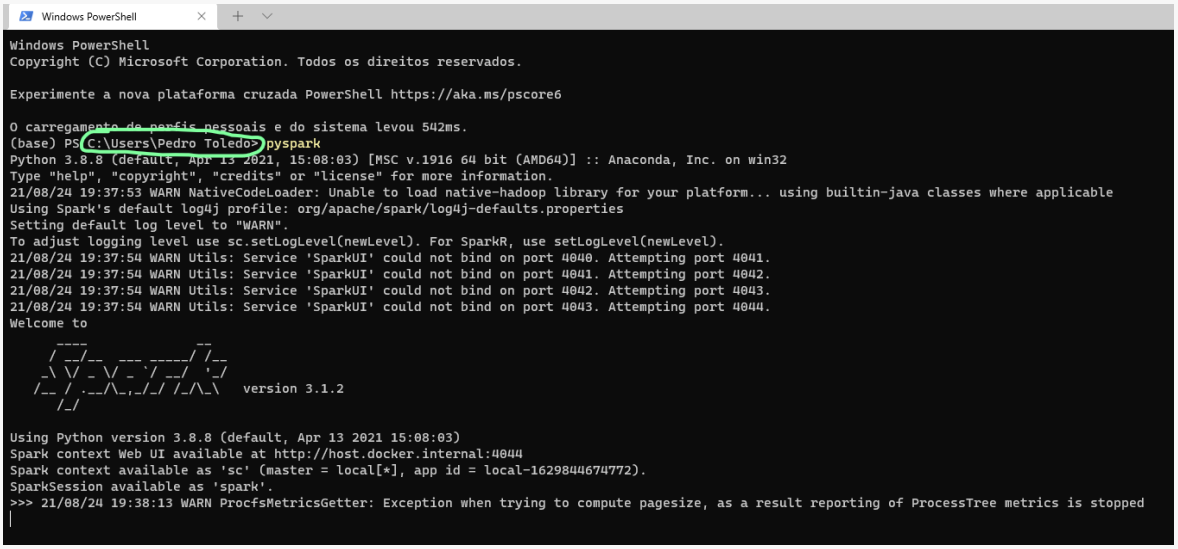
**Obs.:** Faça isso com antecedência, pois as tabelas a serem baixadas são grandes.

**Figura 2 – Download dos Dados**




**Importante:** É necessário que os dados estejam localizados na mesma pasta em que o shell do Spark esteja sendo executado para que eles possam ser lidos de forma mais fácil. Por isso, observe bem onde a shell está sendo executada ou mude o diretório da execução para a mesma pasta dos dados:

**Figura 3 – Identificando o diretório de execução da Shell do Spark**

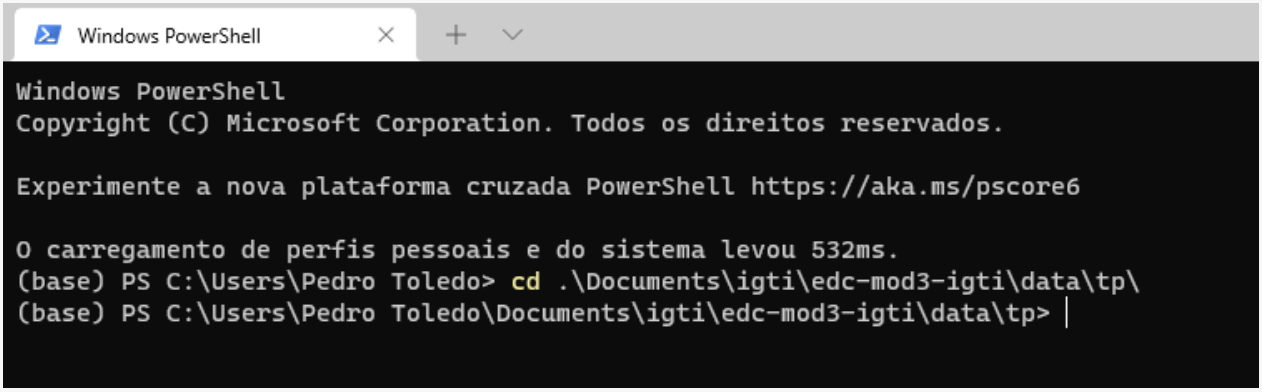



No caso dessa prática, trocaremos o diretório utilizando o comando **cd** (Windows e Linux):

***cd .\Documents\igti\edc-mod3-igti\data\tp***

**Obs.:** O aluno deve utilizar o caminho dos dados no seu próprio computador.

**Figura 4 – Mudando de diretório**




Depois disso, basta executar a shell do Spark novamente.

## 3 – Leitura dos Dados

Após executar novamente a shell do Spark, será realizada a leitura dos dados. Execute os seguintes comandos em sequência:

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

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

A partir desse momento, os dois DataFrames estarão disponíveis para manipulação no ambiente de execução. A prática consiste em manipular os DataFrames de forma a responder algumas das perguntas apresentadas. Além disso, temos algumas perguntas teóricas sobre o Spark e seu funcionamento.

Para um dicionário dos dados sendo trabalhados, acesse:

https://www.imdb.com/interfaces/

## Respostas Finais

Os alunos deverão desenvolver a prática e, depois, responder às seguintes questões objetivas: