

Bootcamp: Engenheiro(a) de Dados (Cloud)**Trabalho Prático**

Módulo 4	Tecnologias de Big Data - Processamento de Dados Massivos
-----------------	--

Objetivos

Exercitar os seguintes conceitos trabalhados no Módulo:

- ✓ Contexto de Big Data e das ferramentas de processamento massivo.
- ✓ Fundamentos do Spark.
- ✓ Funcionamento interno do Spark.
- ✓ 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:

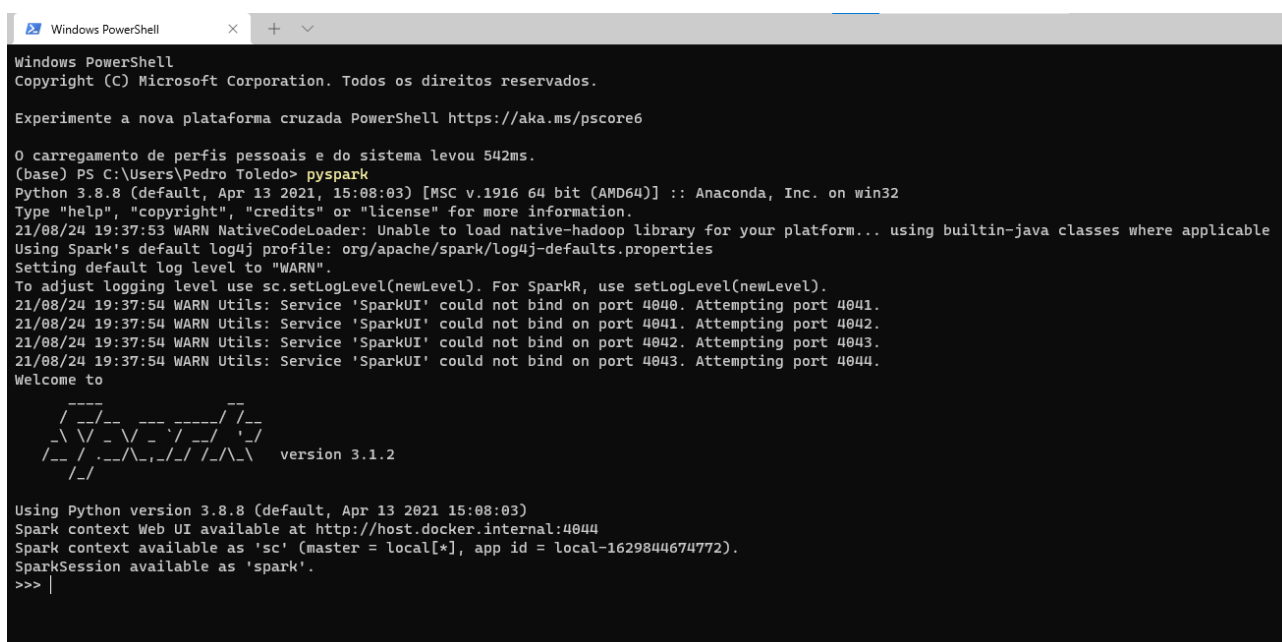
1 – Instalação do Apache Spark

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.



```
Windows PowerShell
Copyright (C) Microsoft Corporation. Todos os direitos reservados.

Experimente a nova plataforma cruzada PowerShell https://aka.ms/pscore6

O carregamento de perfis pessoais e do sistema levou 542ms.
(base) PS C:\Users\Pedro Toledo> pyspark
Python 3.8.8 (default, Apr 13 2021, 15:08:03) [MSC v.1916 64 bit (AMD64)] :: Anaconda, Inc. on win32
Type "help", "copyright", "credits" or "license()" for more information.
21/08/24 19:37:53 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
21/08/24 19:37:54 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
21/08/24 19:37:54 WARN Utils: Service 'SparkUI' could not bind on port 4041. Attempting port 4042.
21/08/24 19:37:54 WARN Utils: Service 'SparkUI' could not bind on port 4042. Attempting port 4043.
21/08/24 19:37:54 WARN Utils: Service 'SparkUI' could not bind on port 4043. Attempting port 4044.
Welcome to

  ____      _
 / ___|  _ \| | | |
 \___ \ | | | | | | |
  ___) || |_| | | | |
 /___ \||___|_|_|_|_|
       |

version 3.1.2

Using Python version 3.8.8 (default, Apr 13 2021 15:08:03)
Spark context Web UI available at http://host.docker.internal:4044
Spark context available as 'sc' (master = local[*], app id = local-1629844674772).
SparkSession available as 'spark'.
>>> |
```

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/1DfuJmIOsXU8hgCRUui-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.

```

Windows PowerShell
Copyright (C) Microsoft Corporation. Todos os direitos reservados.

Experimente a nova plataforma cruzada PowerShell https://aka.ms/pscore6

O carregamento de perfis pessoais e do sistema levou 542ms.
(base) PS C:\Users\Pedro Toledo> pyspark
Python 3.8.8 (default, Apr 13 2021, 15:08:03) [MSC v.1916 64 bit (AMD64)] :: Anaconda, Inc. on win32
Type "help", "copyright", "credits" or "license" for more information.
21/08/24 19:37:53 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
21/08/24 19:37:54 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
21/08/24 19:37:54 WARN Utils: Service 'SparkUI' could not bind on port 4041. Attempting port 4042.
21/08/24 19:37:54 WARN Utils: Service 'SparkUI' could not bind on port 4042. Attempting port 4043.
21/08/24 19:37:54 WARN Utils: Service 'SparkUI' could not bind on port 4043. Attempting port 4044.
Welcome to

  ____
 /  __ \
/   /  \
/_  /    \
/_ /      \
 \  \      \
  \  \__  \
   \___/  \
    \___/

version 3.1.2

Using Python version 3.8.8 (default, Apr 13 2021 15:08:03)
Spark context Web UI available at http://host.docker.internal:4044
Spark context available as 'sc' (master = local[*], app id = local-1629844674772).
SparkSession available as 'spark'.
>>> 21/08/24 19:38:13 WARN ProcsMetricsGetter: Exception when trying to compute pagesize, as a result reporting of ProcessTree metrics is stopped

```

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.

```

Windows PowerShell
Copyright (C) Microsoft Corporation. Todos os direitos reservados.

Experimente a nova plataforma cruzada PowerShell https://aka.ms/pscore6

O carregamento de perfis pessoais e do sistema levou 532ms.
(base) PS C:\Users\Pedro Toledo> cd .\Documents\igti\edc-mod3-igti\data\tp\
(base) PS C:\Users\Pedro Toledo\Documents\igti\edc-mod3-igti\data\tp> |

```

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/>