SEMANA 08 PROGRAMA DE BOLSAS COMPASSO RAFAEL IGNAULIN

1) Instale a última versão do Apache Spark em Python via Pip



Download Libraries → Documentation → Examples Commu

Download Apache Spark™

- 1. Choose a Spark release: 3.1.2 (Jun 01 2021) V
- 2. Choose a package type: Pre-built for Apache Hadoop 3.2 and later v
- 3. Download Spark: spark-3.1.2-bin-hadoop3.2.tgz
- Site do Apache Spark, baixando .tar.gz



- Arquivo comprimido do spark



Descompactado

```
# If you come from bash you might have to change your $PATH.

# export PATH=$HOME/bin:/usr/local/bin:$PATH

# Path to your oh-my-zsh installation.
export ZSH="/home/rafael_ignaulin/.oh-my-zsh"

# Path for JDK 11
export PATH=$PATH:/usr/bin/jdk11

# Path for SPARK

export SPARK_HOME=~/apps/spark-3.1.2-bin-hadoop3.2
export PATH=$PATH:$SPARK_HOME/bin
export PYTHONPATH=$SPARK_HOME/python:$PYTHONPATH
export PYSPARK_PYTHON=python3
export PATH=$PATH:$JAVA_HOME/jre/bin
```

- Arquivo .zshrc , configurado variaveis ambiente para o SPARK
- Comando pyspark funcionando:

```
Welcome to

/ _ / _ / _ _ _ _ _ _ _ _ / _ _ _ _ / _ _ _ _ / _ _ _ _ / _ _ _ _ / _ _ _ _ / _ _ _ _ / _ _ _ _ / _ _ _ _ / _ _ _ _ _ / _ _ / _ _ / _ _ / _ _ / _ _ / _ _ / _ _ / _ _ / _ _ / _ / _ / _ _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _ / _
```

2) Usando o Spark Shell faça um programa que conte as palavras de um arquivo README.md da pasta home do Spark – Somente da Seção Basics do link

```
>>> textFile = spark.read.text("apps/spark-3.1.2-bin-hadoop3.2/README.md")
>>> textFile.count()
108
>>> textFile.first()
Row(value='# Apache Spark')
>>> linesWithSpark = textFile.filter(textFile.value.contains("Spark"))
>>> textFile.filter(textFile.value.contains("Spark"))
```

- Tutorial do Spark, parte Basics.