ESPECIALIZAÇÃO EM CIÊNCIAS DE DADOS - 2ª EDIÇÃO

GERÊNCIA DE INFRAESTRUTURA PARA BIG DATA

PROF. TIAGO FERRETO

US ACCIDENTS DATASET COM SPARK E R

DANIEL KLUG, DANIELLE RODRIGUES GUIDINI, EDUARDO JOSÉ SILVA E MATHEUS DA CONCEIÇÃO EVALDT

US Accidents Dataset

https://www.kaggle.com/sobhanmoosavi/us-accidents?select=US_Accidents_June20.csv

- Cerca de 3,5 milhões de registros de acidentes de trânsito ocorridos entre fevereiro de 2016 e junho de 2020 em 49 estados dos Estados Unidos da América.

- Fontes: departamentos federais e estaduais de transportes, agências policiais, e câmeras e sensores de trânsito espalhados ao longo da malha rodoviária.

- Os dados brutos são disponibilizados em formato tabular, organizados em 49 colunas, e estão armazenados em arquivo com extensão CSV de aproximadamente 1.24GB de tamanho.

Ferramentas

- **Máquina Virtual:** sistema operacional Linux Ubuntu (64-bit), alocação de 4096 MB de memória e disco com alocação dinâmica com 10 GB de tamanho, em formato VDI.
- **Data Storage:** HDFS (Hadoop Distributed File System)
- **Data Processing**: Spark
- **Data Analysis**: RStudio

Instalação Hadoop

- **Shell script para instalação do Hadoop e carregamento do banco de dados para o HDFS:**

  https://github.com/daniellerguidini/spark_bigdata/blob/main/hadoop-config.sh

1. Fazer o download do arquivo *hadoop-config.sh*
2. Execução do seguinte comando, via terminal:

   chmod a+rwx hadoop-config.sh && . ./hadoop-config.sh

Instalação R

- **Shell script para instalação do R e dependências dos pacotes**

  https://github.com/daniellerguidini/spark_bigdata/blob/main/R-config.sh

1. Fazer o download do arquivo *R-config.sh*
2. Execução do seguinte comando, via terminal:

   chmod a+rwx R-config.sh && . ./R-config.sh

Package Manager

- Pacotes do R tentam ser compilado no momento da instalação, portanto, por exemplo, se as dependências C++ não estão instaladas, diversos erros podem aparecer.

- RStudio possui uma plataforma chamada *Package Manager:*

https://packagemanager.rstudio.com/client/#/repos/1/packages

- Nesta plataforma estão disponibilizados todos os pacotes do CRAN com os binários já compilados.

RStudio

1. Acesse https://rstudio.com/products/rstudio/download/#download

2. Realize o download da versão para o sistema operacional Ubuntu 18/Debian 10.

3. Execute o arquivo baixado e siga os passos da instalação padrão.

Integração Spark com RStudio

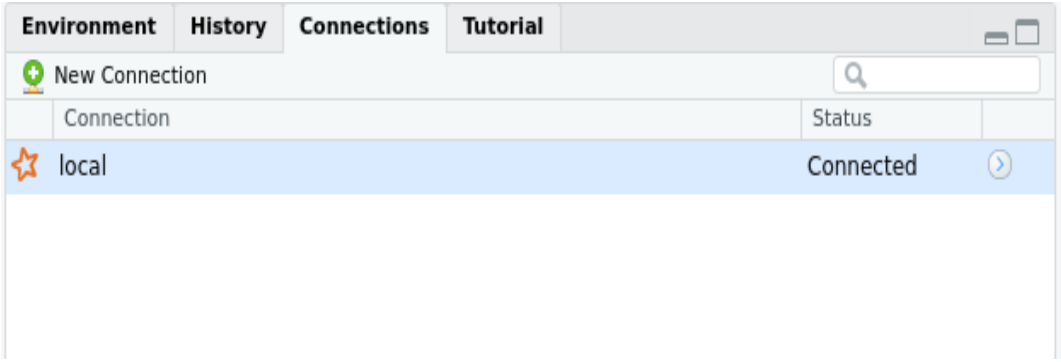
1. Instalar os pacotes *Sparklyr* e *Shiny*
2. Instalar o Spark
3. Conectar com o Spark

```
library(sparklyr)
library(dplyr)

spark_install(version = "1.6.2")
options("scipen"=100, "digits"=8)
sc <- spark_connect(master = "local")
```

Integração Spark com RStudio

1. Instalar os pacotes *Sparklyr* e *Shiny*
2. Instalar o Spark
3. Conectar com o Spark

```
library(sparklyr)
library(dplyr)

spark_install(version = "1.6.2")
options("scipen"=100, "digits"=8)
sc <- spark_connect(master = "local")
```



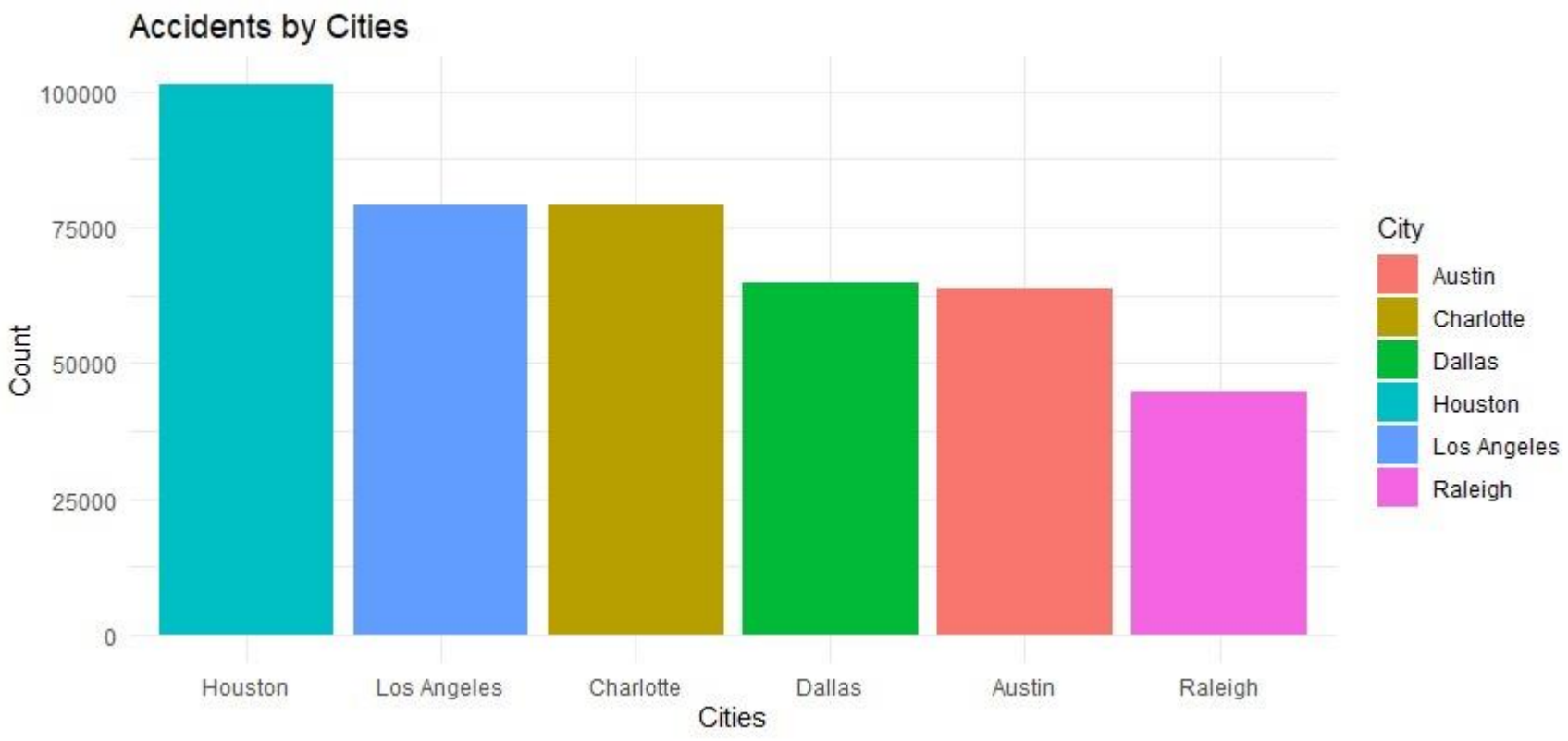
Análise dos dados

- **R script contendo a instalação do Spark pelo Rstudio e as análises**

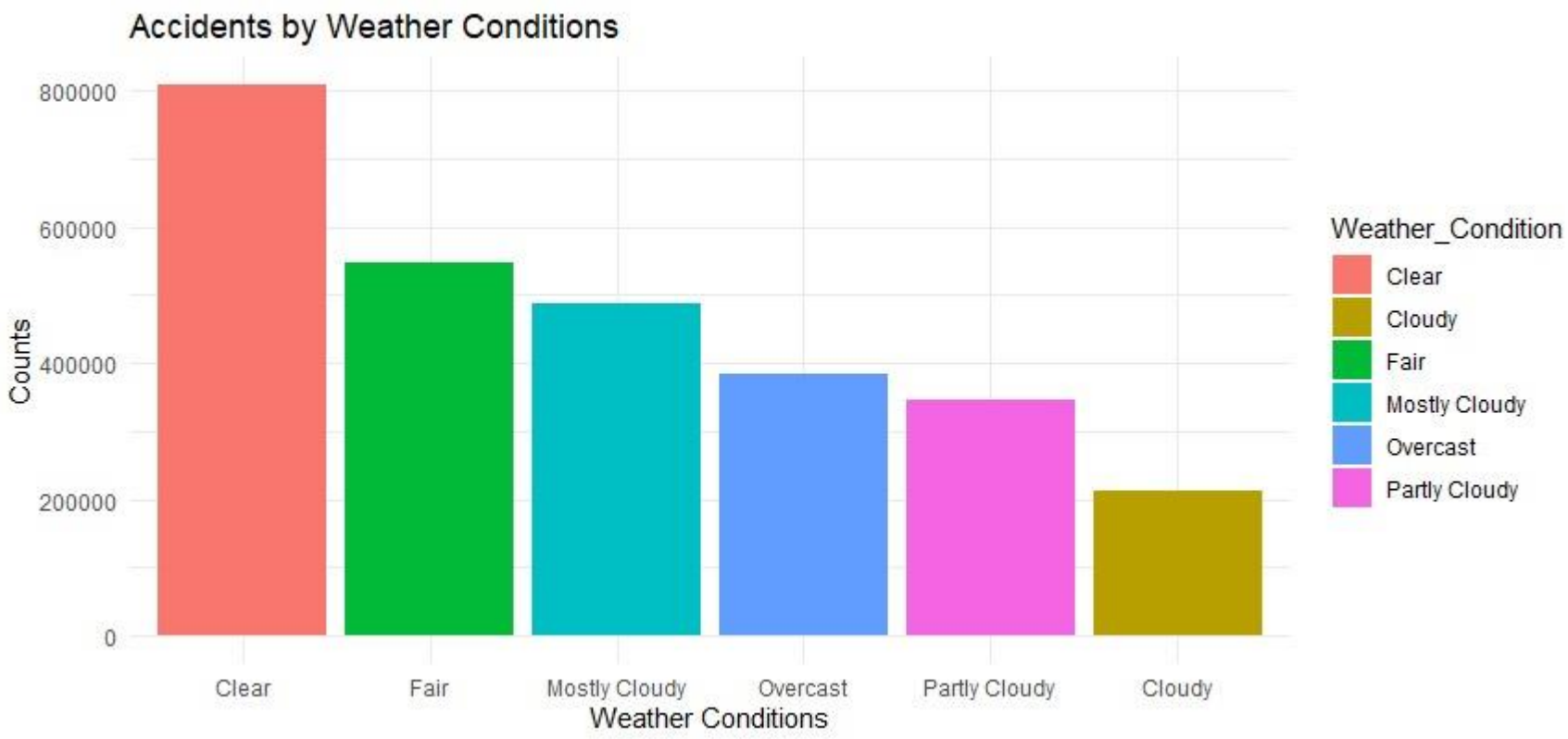
  https://github.com/daniellerguidini/spark_bigdata/blob/main/spark-data.R

- **Perguntas:**
  1. Quais cidades possuem o maior número de acidentes?
  2. Qual a distribuição de acidentes de acordo com as condições climáticas na hora do acidente?
  3. A direção do vento impacta no número de acidentes?

1. Quais cidades possuem o maior número de acidentes?



2. Qual a distribuição de acidentes de acordo com as condições climáticas?



3. A direção do vento impacta no número de acidentes?