Análise do *dataset* US-Accidents: A Countrywide Traffic Accident Dataset com Spark e R

Daniel Klug, Danielle Guidini, Eduardo José Silva, Matheus Evaldt

prof.daniel.klug@gmail.com, daniellerguidini@gmail.com,
eduardo.jose@pucrs.edu.br,mevaldt88@gmail.com

Resumo. *Este artigo descreve uma proposta de solução para a configuração de ambiente de trabalho e posterior implementação de uma solução computacional para processamento e análise de Big Data, utilizando, para tanto, as tecnologias Hadoop, Spark, R e RStudio.*

Abstract. *This paper describes a solution proposal for the configuration of a work environment and subsequent implementation of a computational solution for Big Data processing and analysis, using the technologies Hadoop, Spark, R and RStudio.*

1. Introdução

No contexto da disciplina de Gerência de Infraestrutura para Big Data, da 2ª edição do curso de pós-Graduação em Ciência de Dados da PUCRS, foi proposta a realização de um projeto em grupo visando colocar em prática as noções conceituais abordadas ao longo do período letivo. A primeira etapa da tarefa consistiu em escolher um conjunto de dados de interesse, recaindo a escolha sobre a versão 3 do *dataset US-Accidents: A Countrywide Traffic Accident Dataset*, encontrado por meio da plataforma *Kaggle*, onde é disponibilizado de forma gratuita para finalidade não comercial, de pesquisa ou acadêmica sob a licença *Creative Commons Attribution-Noncommercial-ShareAlike license* (CC BY-NC-SA 4.0).

Uma vez obtidos os dados, configuramos um ambiente de Big Data local com o uso da ferramenta Hadoop. Ato contínuo, as tecnologias R e Spark foram integradas na *IDE* RStudio e, então, empregadas para realizar a ingestão do *dataset* no ambiente de Big Data e para conduzir a análise dos dados e responder a um conjunto definido de perguntas formuladas sobre seu domínio.

Por fim, em acréscimo à solução computacional desenvolvida, foi elaborado o presente relatório descritivo, que é complementado por uma apresentação tele presencial em que os autores fornecem um *overview* do desenvolvimento do projeto e dos resultados obtidos. A íntegra do projeto, incluindo os *script* de configuração do ambiente e os arquivos fonte em R das análises conduzidas sobre o *dataset,* está armazenada em repositório aberto do GitHub, podendo ser acessado no endereço https://github.com/daniellerguidini/spark_bigdata/blob/main/spark-data.R

2. Descrição do *dataset*

O *dataset* selecionado agrupa 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. Os dados nele contidos foram coletados de diversas fontes, tais como os

departamentos federais e estaduais de transportes, agências policiais, e câmeras e sensores de trânsito espalhados ao longo da malha rodoviária[1].
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.

3. Configuração do ambiente

A aplicação desenvolvida executa sobre uma máquina virtual (VM) Linux Ubuntu (64-bit), criada com o auxílio da ferramenta Oracle VM Virtual Box.
Como sistema operacional para big data foi utilizada a ferramenta Hadoop. Por meio do sistema de arquivos distribuídos (HDSF) e do gerenciador de carga de trabalho e de recursos (YARN), seus principais componentes, o Hadoop é capaz de distribuir o processamento analítico da aplicação entre várias máquinas, cada uma delas trabalhando simultaneamente em suas próprias porções de dados. Com isso, o Hadoop consegue atender os requisitos desejados de tolerância à falhas, de capacidade de recuperação, de consistência e de escalabilidade[2]. No contexto de nossa aplicação o Hadoop foi utilizado para realizar a ingestão dos dados.
A aplicação foi codificada na linguagem R e desenvolvida na IDE RStudio, visando superar a inexistência de interface direta da linguagem para o Linux. Considerando a necessidade de processamento de *big data* foi acoplado o pacote Sparklyr, que provê uma interface para a ferramenta Spark.
O Spark consiste em uma abstração de computação distribuída de propósito geral, com foco no processamento de dados, que possibilita alto desempenho nesta espécie de tarefas devido ao uso um modelo de dados chamado RDDs (Conjunto de Dados Distribuídos Resilientes). Os dados são mantidos em memória enquanto são processados, eliminando a necessidade de escritas intermediárias. Com isso, o Spark tira proveito da execução de DAG (Grafo Direcionado Acíclico), otimizando o processamento[3].
Por fim, adicionamos aos recursos nativos da linguagem o pacote Shiny, por meio do qual se tornou possível a construção interativa de aplicativos web diretamente a partir do código escrito em R.
A seguir, descrevemos passo a passo as etapas de configuração do ambiente.

3.1 Instalação da Máquina Virtual

A VM foi configurada através da ferramenta Oracle VM Virtual Box, com as seguintes especificações: 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.

3.2 Instalação do Hadoop

A fim de unificar os passos de instalação do Hadoop e de ingestão dos dados no HDFS, criamos um *shell script* automatizando a execução dos comandos necessários. A seguir, descrevemos a sequência de passos:

---

[1] Adaptado de Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. (2019). Disponível em https://smoosavi.org/datasets/us_accidents. Acesso em 15/12/2020.
[2] Bengfort, Benjamin. KIM, Jenny (2016), p.34-35.
[3] *Ibidem*, p.100

1. Fazer o download do arquivo *hadoop-config.sh*. O arquivo está disponível no *link* https://github.com/daniellerguidini/spark_bigdata/blob/main/hadoop-config.sh e deve ser salvo na pasta HOME.
2. Execução do seguinte comando, via terminal:

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

3.3 Instalação do R e RStudio

Da mesma forma, visando assegurar a correta instalação do R e de todas as dependências necessárias criamos um *shell script* específico para essa tarefa. Logo, para realizar a instalação do R e a importação dos pacotes, os seguintes passos devem ser executados:
1. Download do arquivo *R-config.sh*. O arquivo está disponível no *link* https://github.com/daniellerguidini/spark_bigdata/blob/main/R-config.sh e deve ser salvo na pasta HOME.
2. Execução do seguinte comando, via terminal:

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

Ao final da execução do *script*, o R e as dependências dos pacotes já estarão instalados. Após, é necessário dar seguimento com a instalação da *IDE* RStudio. Para a realização do download, são necessários os seguintes passos:

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.

Cabe registrar que muitas vezes em ambientes Linux os problemas de instalação envolvendo pacotes de R decorrem da circunstância de que é feita uma tentativa imediata de compilar o pacote. Assim, exemplificativamente, se as dependências de C++ não tiverem sido previamente instaladas, um erro pode surgir.
Essas dificuldades podem ser superadas utilizando-se a ferramenta *Package Manager* do RStudio, que disponibiliza todos os pacotes do CRAN com os binários já compilados, e pode ser obtida no link
https://packagemanager.rstudio.com/client/#/repos/1/packages.
Após a instalação do *Package Manager* deve-se clicar na opção Setup, depois em Binary e, então, escolher o sistema operacional desejado. Retiramos deste local os comandos de instalação das dependências dos pacotes necessários para a nossa aplicação.

3.4 Integração do Spark com o RStudio

A última etapa de configuração do ambiente consiste em integrar o Spark com o R através da *IDE* RStudio. Para isso, seguimos os seguintes passos, no terminal do RStudio:
1. Instalar os pacotes *Sparklyr* e *Shiny:*

```
library(sparklyr)
library(dplyr)
```

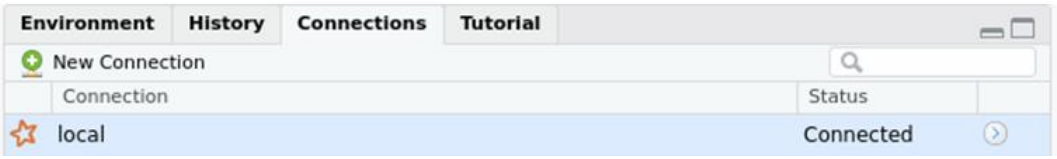
2. Instalar o Spark:

```
spark_install(version = "1.6.2")
options("scipen"=100, "digits"=8)
```

3. Conectar com o Spark:

```
sc <- spark_connect(master = "local")
```

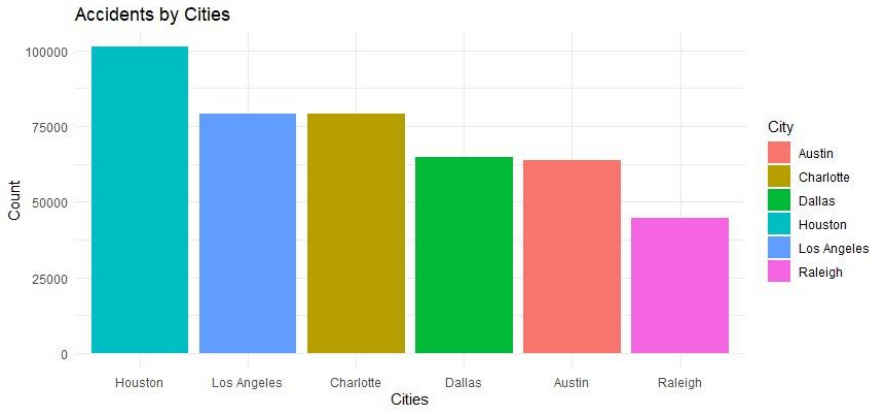
4. Verificar o status da conexão no painel superior à direita:



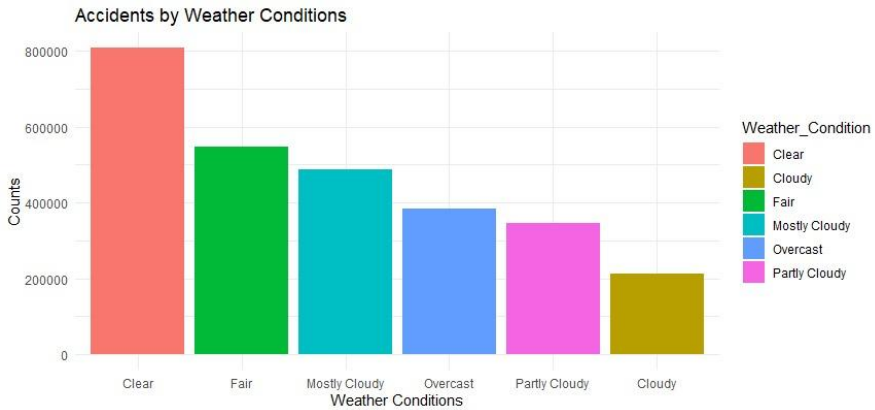
4. Análise do Banco de Dados

Para análise do *dataset* foram levantadas três questões, cujos resultados são apresentados abaixo na forma de gráficos:

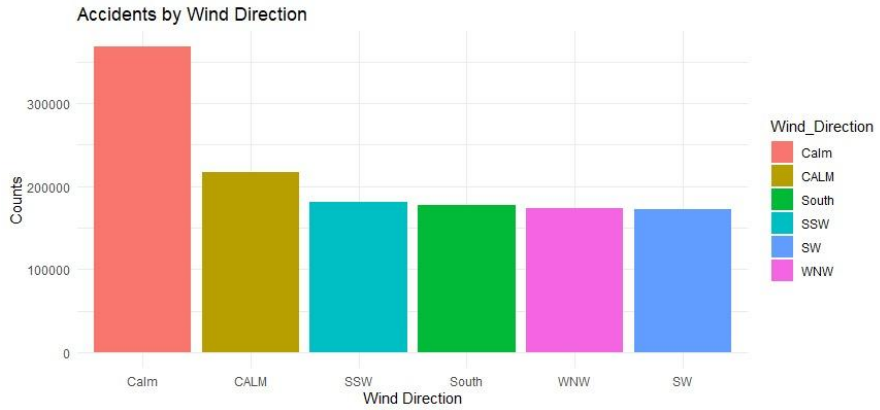
1. Quais cidades possuem o maior número de acidentes?



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?



5. Conclusão

Este projeto demonstra uma alternativa de ambiente para processamento e análise de Big Data, envolvendo desde a escolha do *dataset* e das tecnologias a serem adotadas, da instalação e configuração das ferramentas e de utilização prática. Entendemos que os resultados obtidos demonstram que a aplicação executa com sucesso, possibilitando o processamento e análise de Big Data, denotando o atingimento dos objetivos iniciais. Ademais, há espaço para que a aplicação continue a ser desenvolvida, com a incorporação de novas funcionalidades, de que são exemplo a carga e o processamento dinâmicos de dados, que não exploramos.

Em arremate, embora a escolha de projeto tenha sido pela implementação da solução em ambiente local e com utilização de apenas uma máquina, julgamos que com moderado esforço de adaptação sua utilização também seria possível em ambiente de nuvem e/ou com o uso concorrente de mais de máquina. Nestas hipóteses, estimamos que as ferramentas utilizadas, por suas características, potencializariam os ganhos de desempenho da aplicação no processamento de Big Data.

Referências

Bengfort, Benjamin. KIM, Jenny (2016). Analítica de dados com Hadoop: uma indrodução para cientistas de dados. Novatec, São Paulo.

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. (2019) “A Countrywide Traffic Accident Dataset.”

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath (2019) "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In: Proceedings of the 27th ACM SIGSPATIAL. International Conference on Advances in Geographic Information Systems, ACM.