

Desmistificando Big Data: é possível manipular grandes bases de dados em R?

Samuel Macêdo

Definindo Big Data

- Volume
- Velocidade
- Variedade
- Veracidade
- Valor

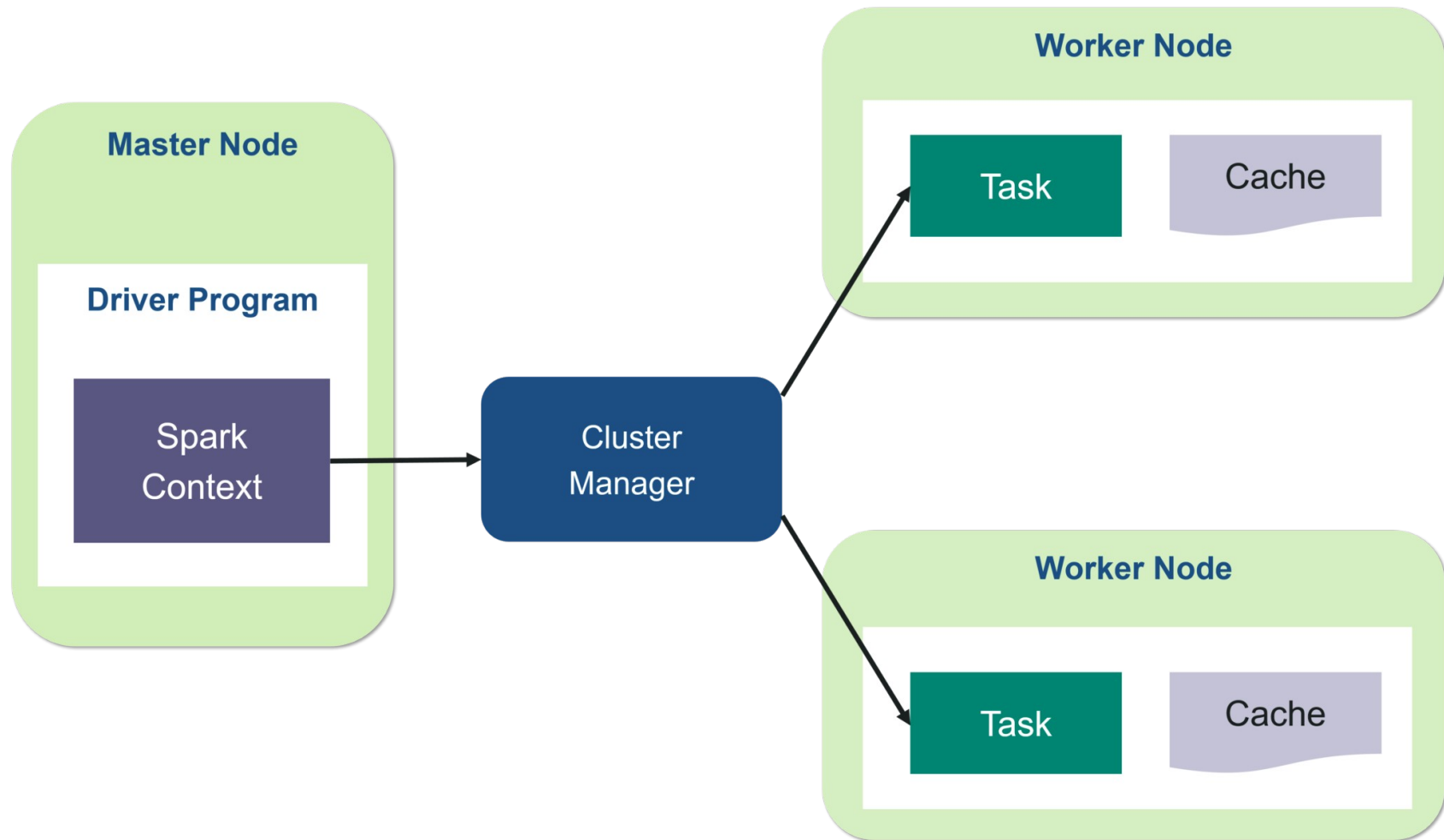
Definindo Big Data

- Veracidade
- Valor
- Volume

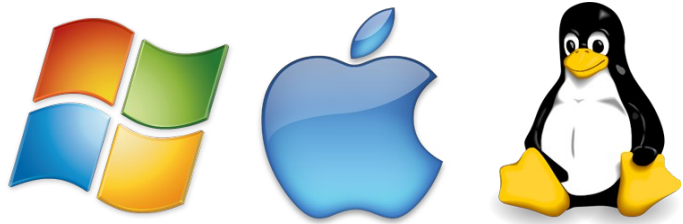
Quem processa Big Data?



Arquitetura



Como funciona o R?



R roda Big Data?

Claro que NÃO

É possível trabalhar com Big Data em R?

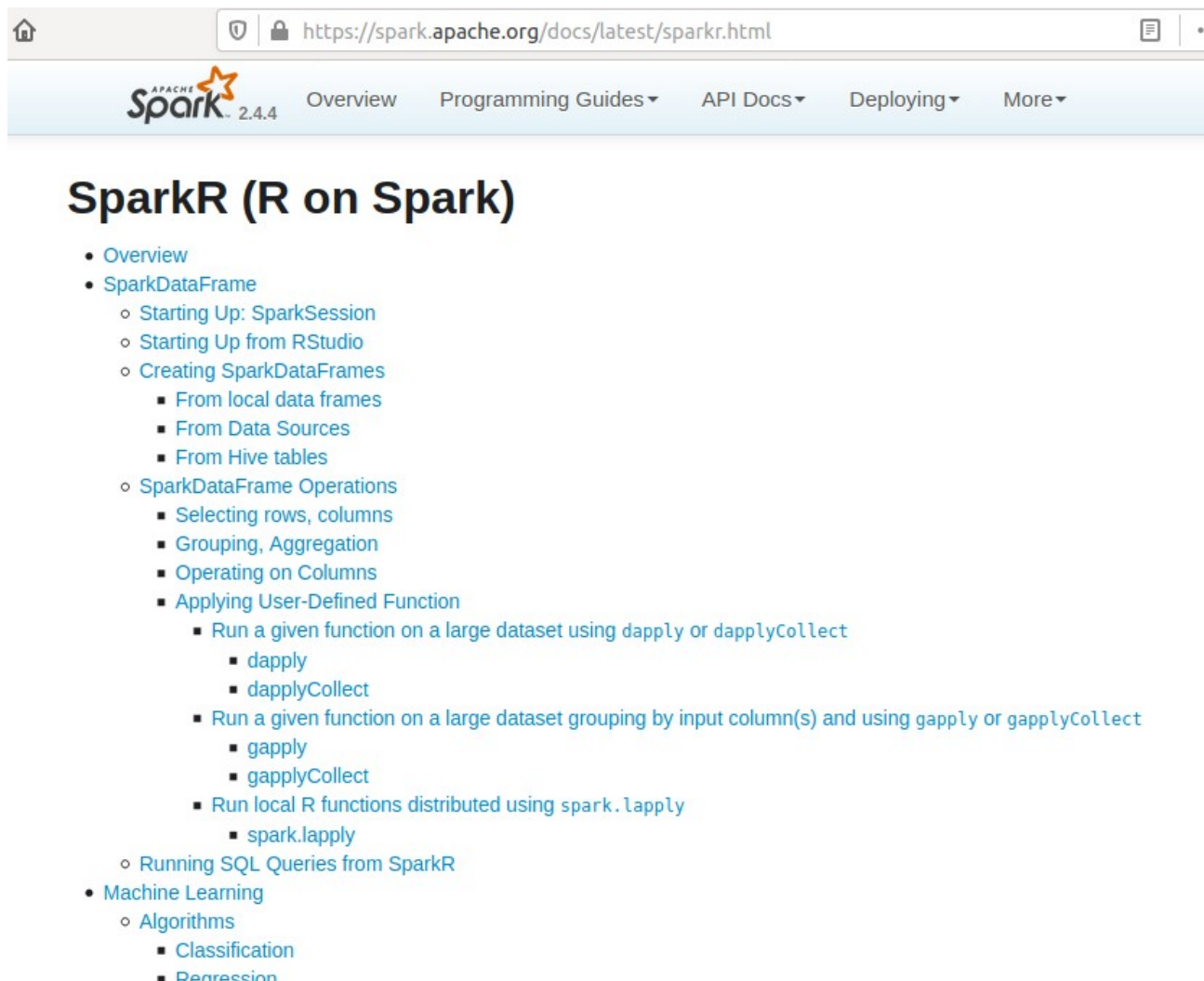
**Definitivamente
SIM**

E qual mágica?



Pacotes em R

- SparkR



The screenshot shows the SparkR documentation page on the Spark website. The browser address bar displays <https://spark.apache.org/docs/latest/sparkr.html>. The navigation bar includes the Spark logo (version 2.4.4) and links for Overview, Programming Guides, API Docs, Deploying, and More. The main heading is "SparkR (R on Spark)". Below it, a table of contents lists the following sections:

- Overview
- SparkDataFrame
 - Starting Up: SparkSession
 - Starting Up from RStudio
 - Creating SparkDataFrames
 - From local data frames
 - From Data Sources
 - From Hive tables
 - SparkDataFrame Operations
 - Selecting rows, columns
 - Grouping, Aggregation
 - Operating on Columns
 - Applying User-Defined Function
 - Run a given function on a large dataset using `dapply` or `dapplyCollect`
 - `dapply`
 - `dapplyCollect`
 - Run a given function on a large dataset grouping by input column(s) and using `gapply` or `gapplyCollect`
 - `gapply`
 - `gapplyCollect`
 - Run local R functions distributed using `spark.lapply`
 - `spark.lapply`
 - Running SQL Queries from SparkR
 - Machine Learning
 - Algorithms
 - Classification
 - Regression

Pacotes em R

- sparklyr

The screenshot shows the website for sparklyr, an R interface for Apache Spark. The page is titled "sparklyr: R interface for Apache Spark". On the left, there is a navigation menu with sections: "Using sparklyr" (containing "Configuring connections" and "Troubleshooting"), "Guides" (containing "Manipulating data", "Machine Learning", "Understanding Caching", "Deployment Options", "Distributed R", "Data Lakes", "ML Pipelines", "Text mining", and "Stream Analysis"), and "sparklyr from RStudio". The main content area features an "Announcement" about a new book, "Mastering Spark with R", available online and in print, with a link to therinspark.com. Below the announcement, there are badges for "build passing", "CRAN 1.0.5", "codecov 80%", "chat", and "on gitter". A list of features is provided: connecting to Spark from R, filtering and aggregating Spark datasets, using Spark's distributed machine learning library, and creating extensions that call the full Spark API. On the right, there is a large orange graphic with the sparklyr logo and a row of buttons for "dplyr", "GraphX (graphframes)", "Streaming", "ML", and "Extensions". The Apache Spark logo is also visible at the bottom of the graphic.

sparklyr from RStudio

dplyr MLib Extensions Streaming News Reference Blog

sparklyr: R interface for Apache Spark

Announcement
The new book is now available on-line and in print!
Visit: therinspark.com for more info

build passing CRAN 1.0.5 codecov 80% chat on gitter

- Connect to [Spark](#) from R. The sparklyr package provides a complete [dplyr](#) backend.
- Filter and aggregate Spark datasets then bring them into R for analysis and visualization.
- Use Spark's distributed [machine learning](#) library from R.
- Create [extensions](#) that call the full Spark API and provide interfaces to Spark packages.

sparklyr

dplyr GraphX (graphframes) Streaming ML Extensions

APACHE SPARK

Invoke

`\\ scala`

`package au.csiro.variantspark.api`

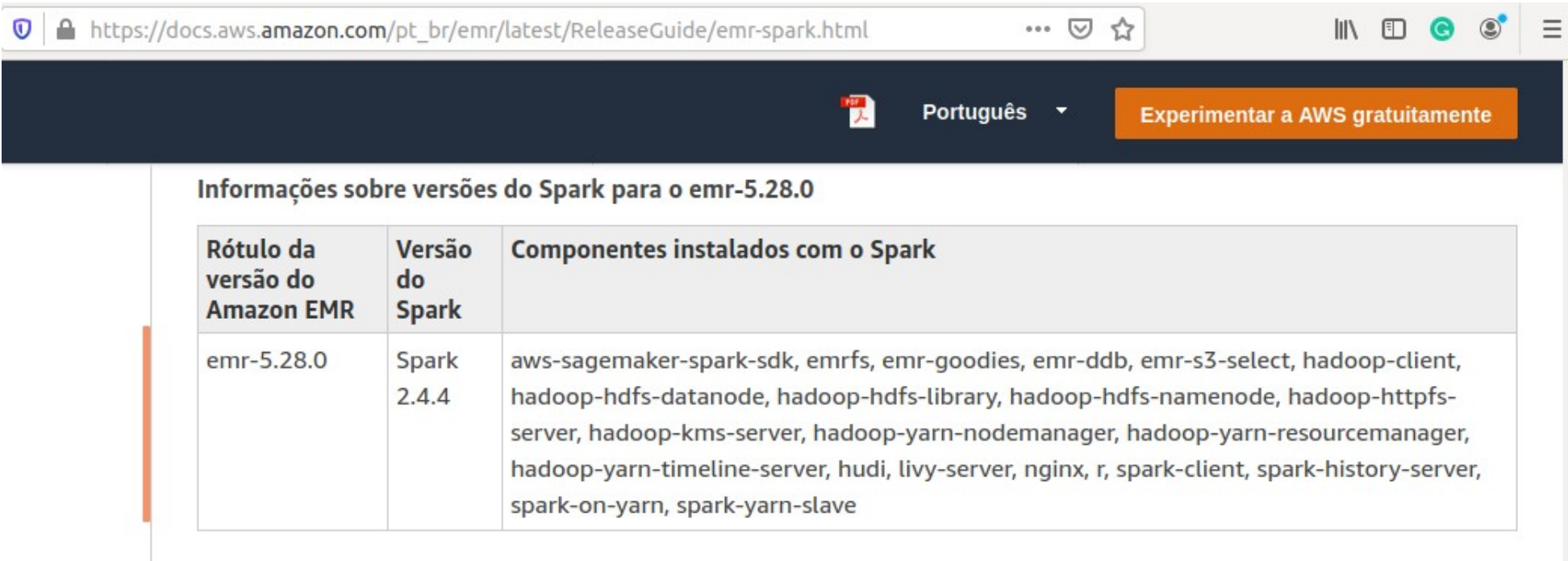
`class VSContext(val spark:SparkSession) {...}`

`\\ r`

**`sparklyr::invoke_new(sc,
"au.csiro.variantspark.api.VSContext", spark_session(sc))`**

Configuração

sparklyr::spark_install()



The screenshot shows a web browser window with the URL https://docs.aws.amazon.com/pt_br/emr/latest/ReleaseGuide/emr-spark.html. The page is in Portuguese and features a dark blue header with a PDF icon, the language 'Português', and a button 'Experimentar a AWS gratuitamente'. The main content area is titled 'Informações sobre versões do Spark para o emr-5.28.0' and contains a table with the following data:

Rótulo da versão do Amazon EMR	Versão do Spark	Componentes instalados com o Spark
emr-5.28.0	Spark 2.4.4	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Eu preciso do R?



JVM
(Java Virtual Machine)



E porque R?

- **Fácil**
- **Concentrado**
- **Aprender uma linguagem**

Quando eu devo usar spark?

- **Use o R sempre que possível!!!**
- **Spark em versão local**
- **Spark em cloud**

Resumo

- **Trate bem seus dados**
- **R como interface para o spark**
- **Só use spark quando realmente precisar**

- **E o principal...**

É possível trabalhar em Big Data com R !!!

Contatos

- samuelmacedo@recife.ifpe.edu.br
- github.com/samuelmacedo83