

Spark

- NPL com Spark
- Utilizando bibliotecas Nativas de ML do Spark
- Classificação: Spam





Como Utilizar Spark?

- Instalar a versão Open Source (Windows, Mac, Linux)
 - <https://spark.apache.org/>
- Instalar no Google Colab
- Utilizar um provedor na Nuvem
 - AWS (EMR)
 - Azure (HDInsight)
 - Databricks
 - Você pode instalar em um servidor na nuvem

Databricks



- Dos criadores do Spark
- Community Edition: sem custo, com algumas limitações
- <https://community.cloud.databricks.com/login.html>



Sign In to Databricks Community Edition



fernando@evoluth.com.br



.....

[Forgot Password?](#)

Sign In

New to Databricks? [Sign Up.](#)

[Privacy Policy](#) | [Terms of Use](#)

Please tell us about yourself

First Name: *

Last Name: *

Company *

Company Email *

Title *

Phone Number

☐ Keep me informed with occasional updates about Databricks
and related open source products

By Clicking "Get Started For Free", you agree to the [Privacy Policy](#).

GET STARTED FOR FREE

O que é Spark?



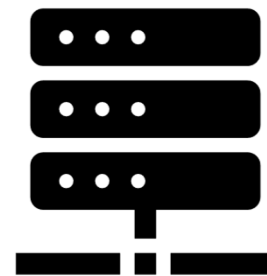
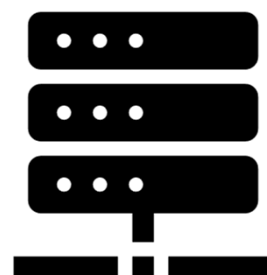
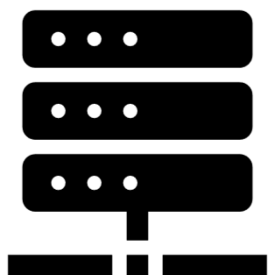
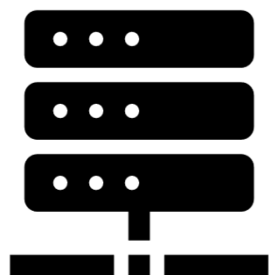
The image features a dark blue background on the left side, which contains a large, lighter blue circular graphic. The word "Spark" is written in white, sans-serif font, positioned over the circular graphic.

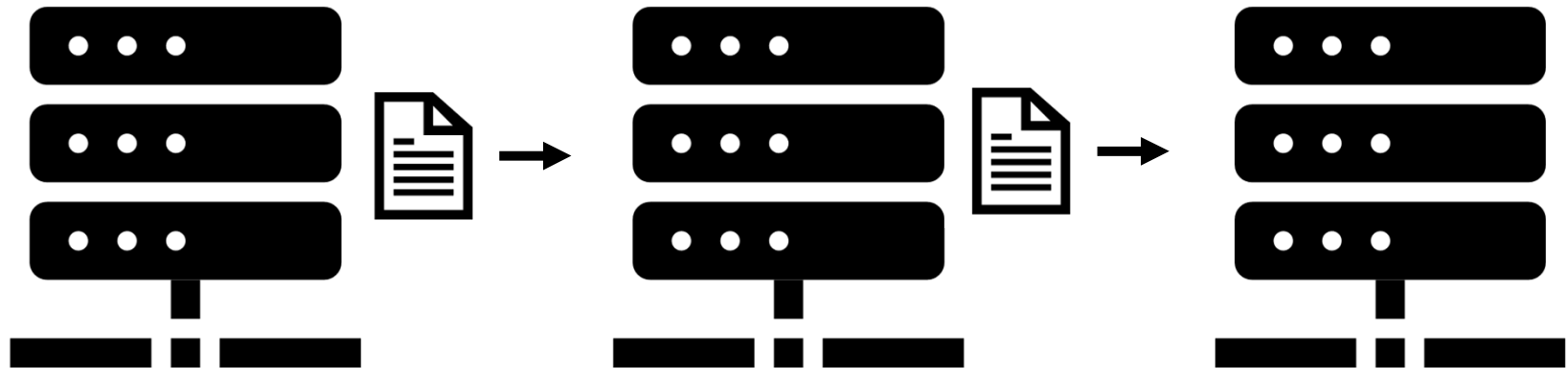
Spark

- Ferramenta de Processamento de Dados Distribuído em um Cluster
- Em memória
- Veloz
- Escalável
- Particionamento

Spark

- Escala horizontalmente - Cluster

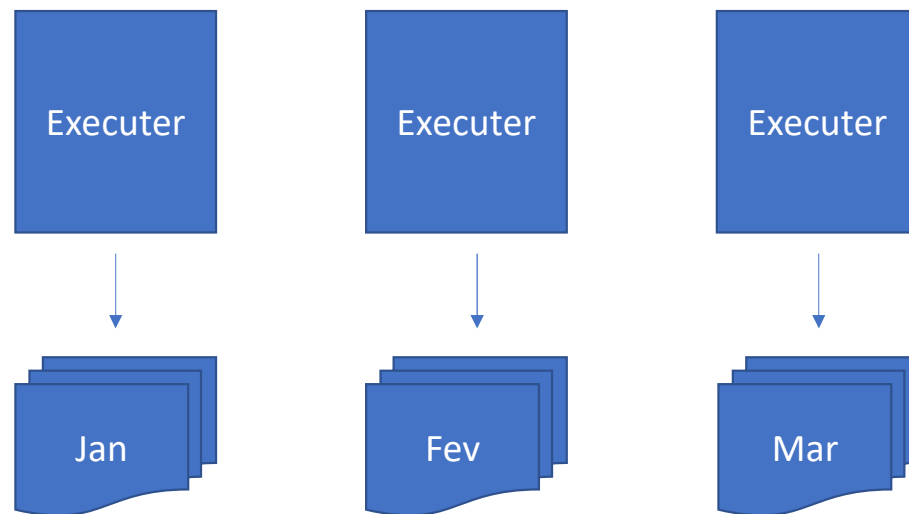




Replicação / Tolerância a Falha

- Dados são copiados entre os nós do cluster. Isso traz o benefício de, entre outras coisas, tolerância a falhas

Particionamento



Spark VS Python, R ou Banco de Dados

- Você precisa Processar dados!
- Custo computacional: CPU, Memória, Rede etc.
- Spark tem arquitetura voltada a processar dados!
 - Melhor performance, porém:
 - Não substitui Python
 - Não substitui SQL ou um SGBDR

Linguagens

Scala 

Python 

Java 

R 

SQL 



Por que Spark?

- NLP são tarefas com alto custo computacional
- Spark Alta performance pela sua natureza “distribuída”
- Com Pyspark, você tem tudo do Python + Spark!

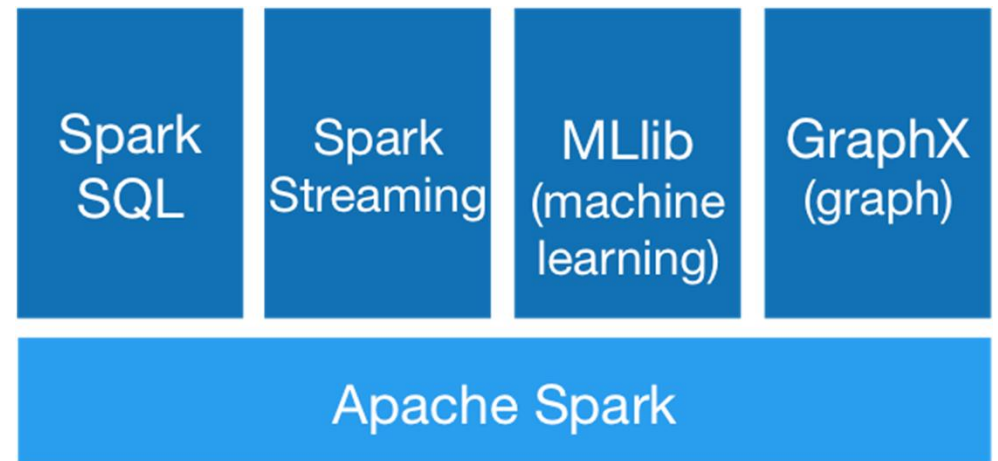


Arquitetura e Componentes



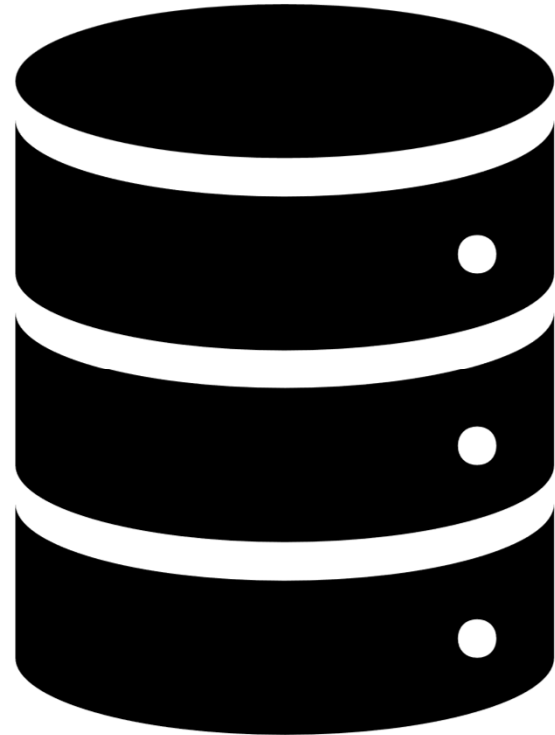
Componentes

- Machine Learning (Mlib)
- SQL (Spark SQL)
- Processamento em Streaming
- Processamento de Grafos (GraphX)



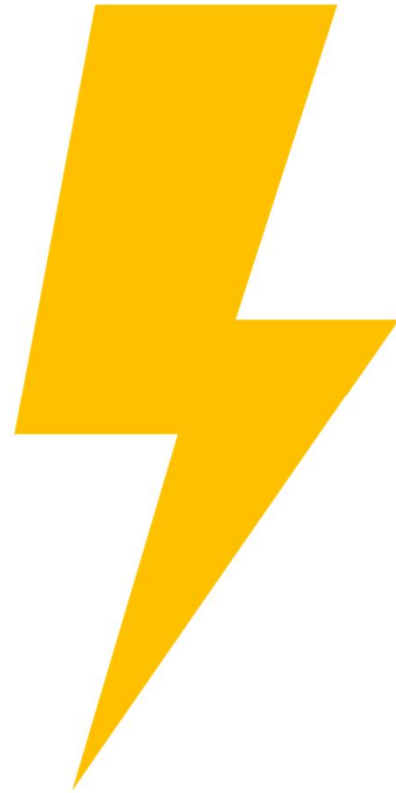
Spark SQL

- Permite ler dados tabulares de várias fontes (CSV, Json, Parquet, ORC etc)
- Pode usar sintaxe SQL



Streaming: Spark Structured Streaming

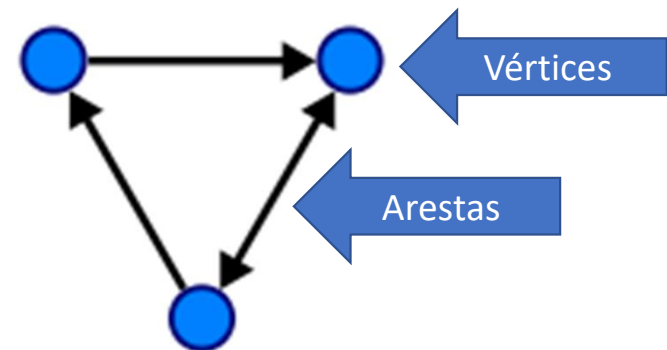
- Dados estruturados





Grafos acíclicos dirigidos

- Spark Constrói Gráficos Acíclicos Dirigidos

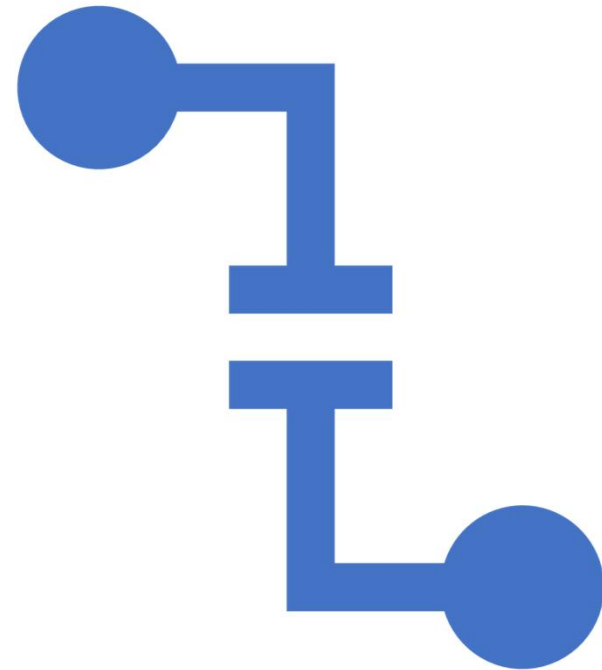


Elementos

- SparkSession: Seção
- Application: Programa

Transformações e Ações

- Um data frame é imutável: traz tolerância a falha
- Uma transformação gera um novo data frame.
- O processamento de transformação de fato só ocorre quando há uma Ação: Lazy Evaluation



Lazy Evaluation

