Spark

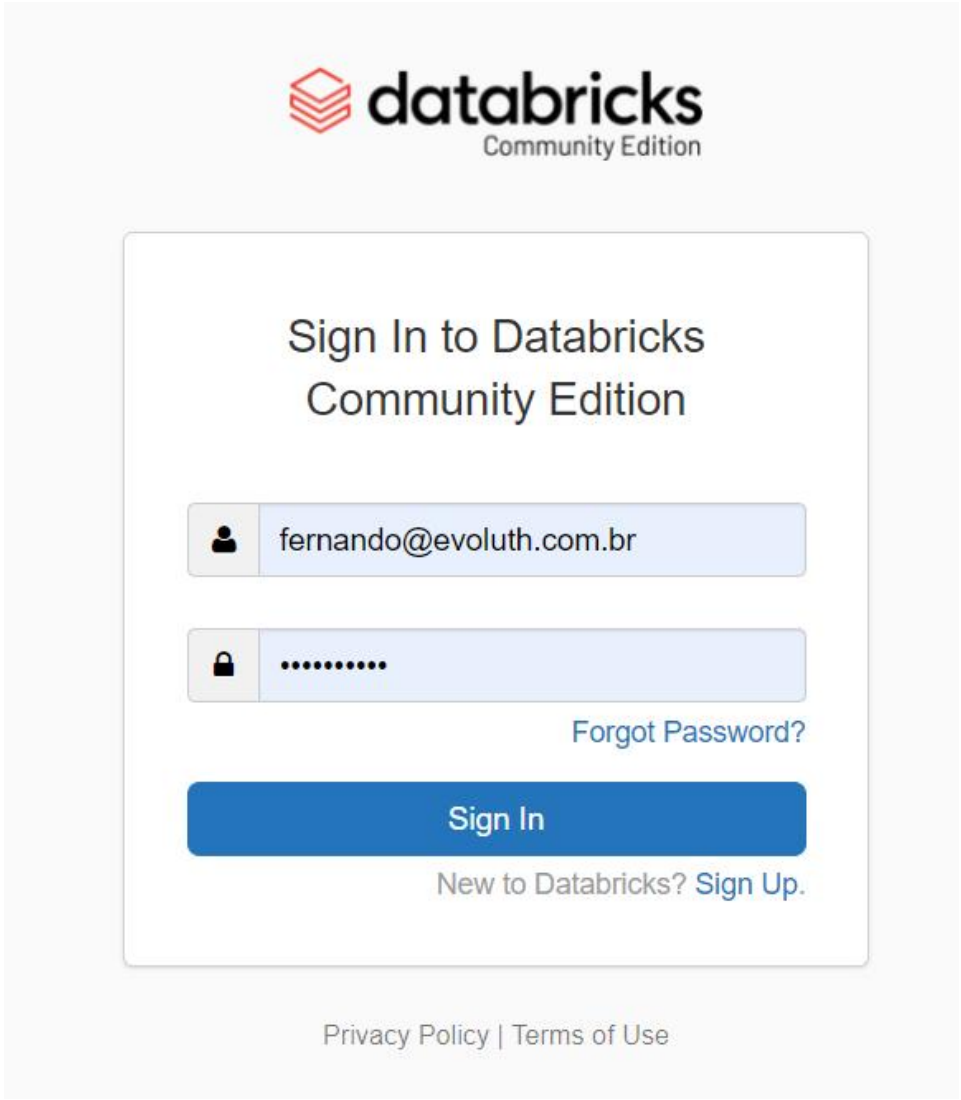
- NPL com Spark
- Utilizando bibliotecas Nativas de ML do Spark
- Classificação: Spam

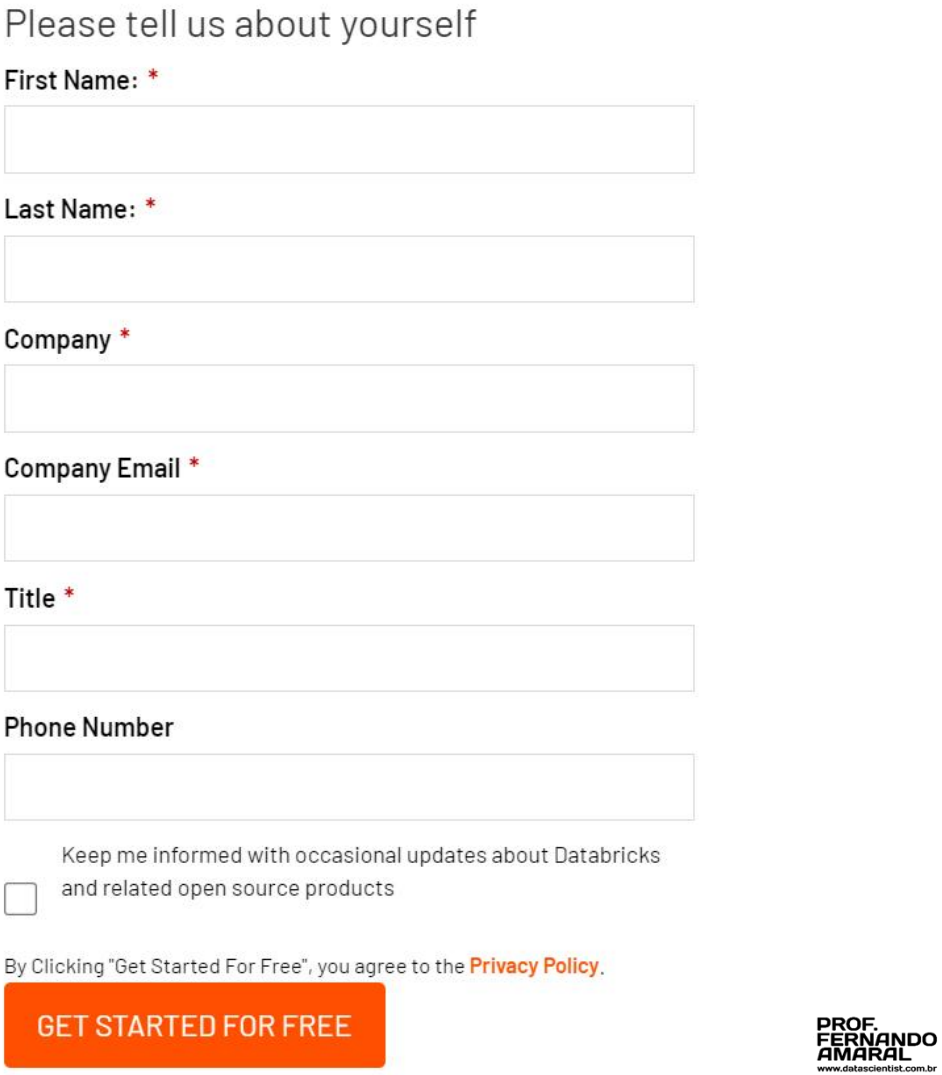
Como Utilizar Spark?

- 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





PROF. FERNANDO AMARAL
www.datascientist.com.br

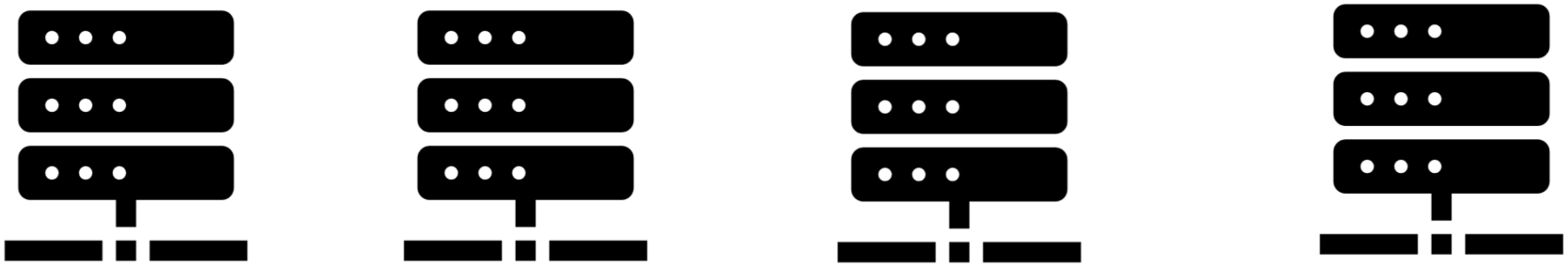
O que é Spark?

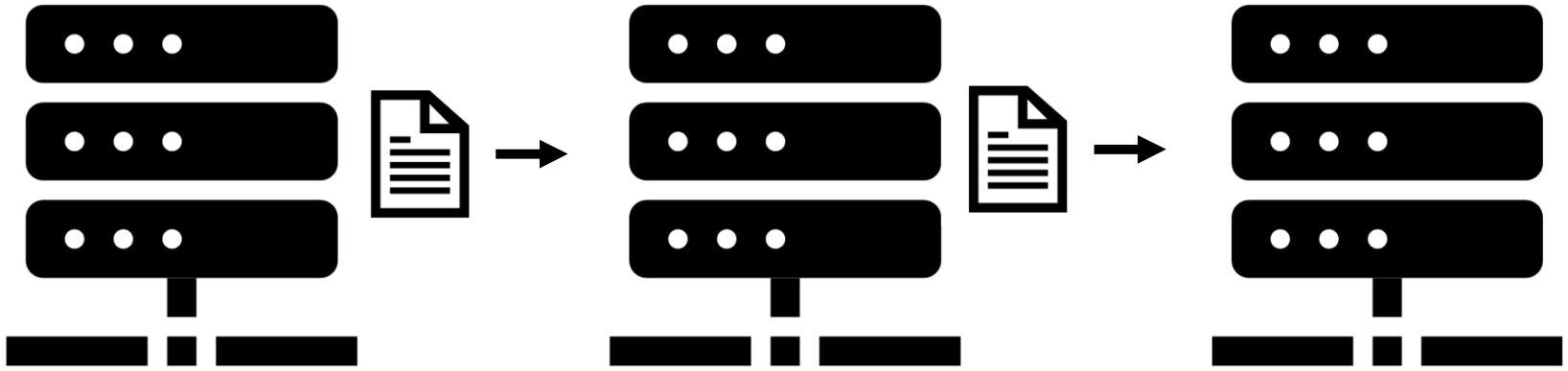
Spark

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

Spark

- Escala horizontalmente - Cluster

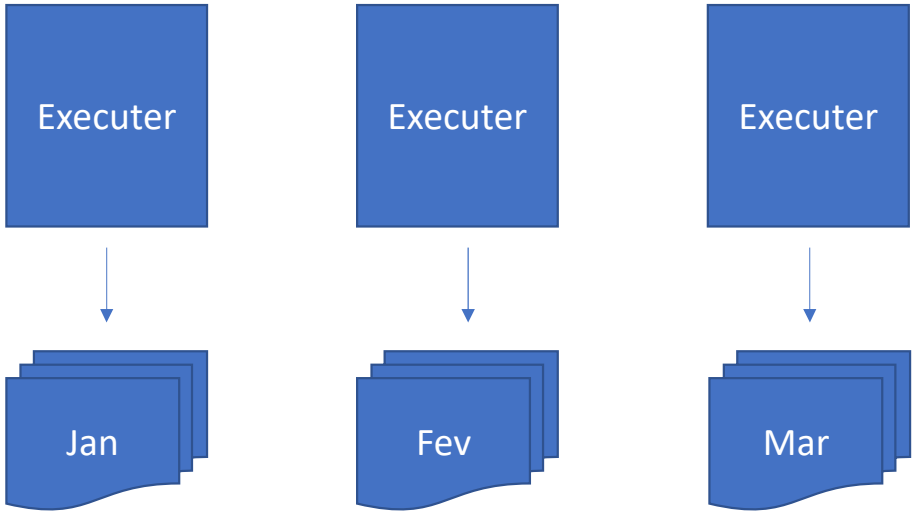




Replicação / Tolerância a Falha

- 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?

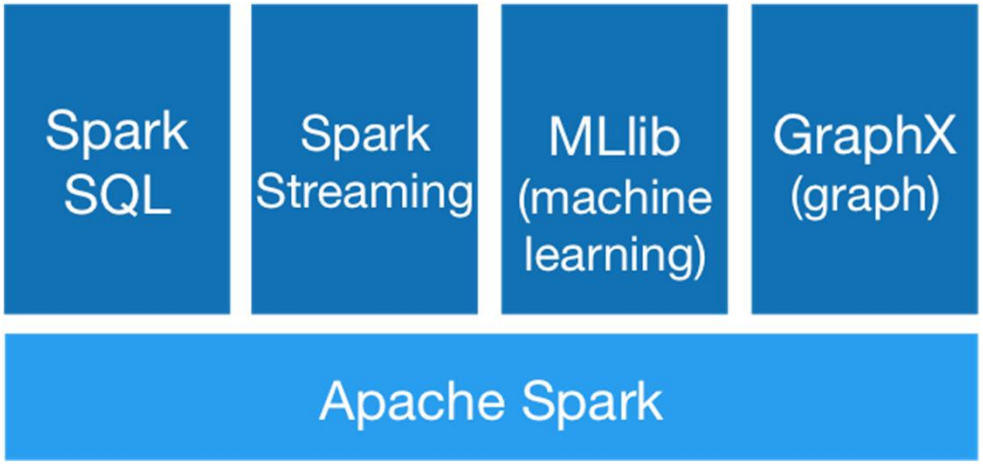
- NLP são tarefas com alto custo computacional
- Spark Alta performance pela sua natureza “distribuída”

- Com Pyspark, você tem tudo do Python + Spark!

Arquitetura e Componentes

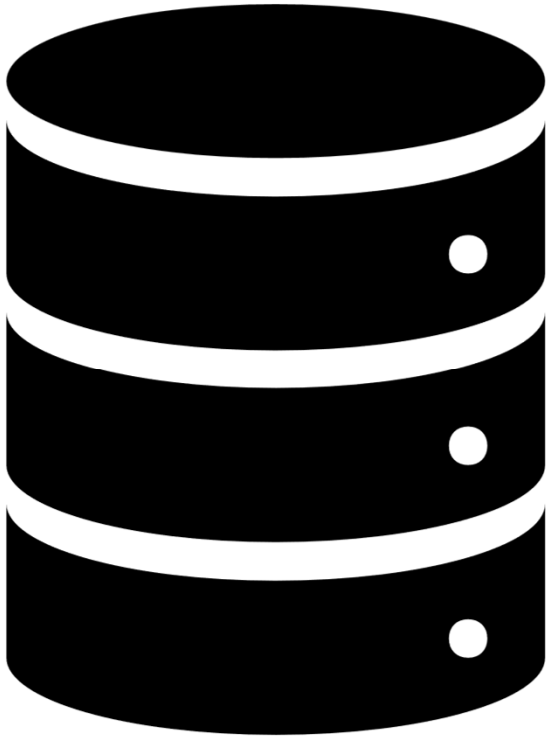
Componentes

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

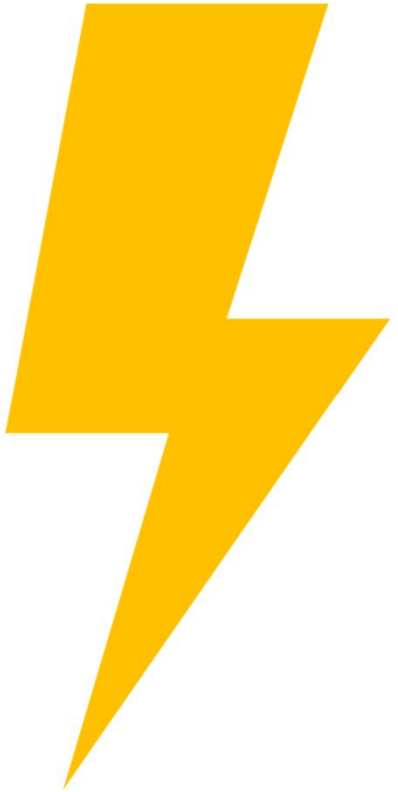


Spark SQL

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



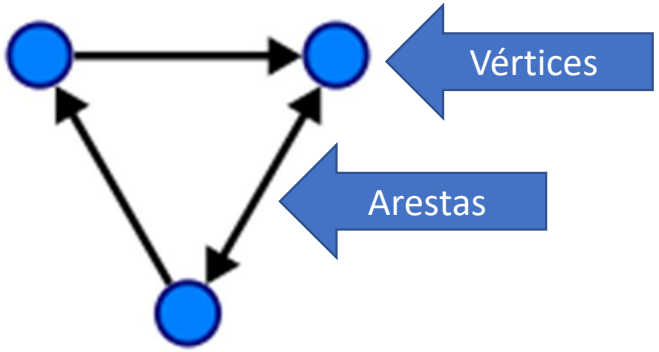
Streaming: Spark Structured Streaming



- Dados estruturados

Grafos acíclicos dirigidos

- Spark Constrói Gráficos Acíclicos Dirigidos



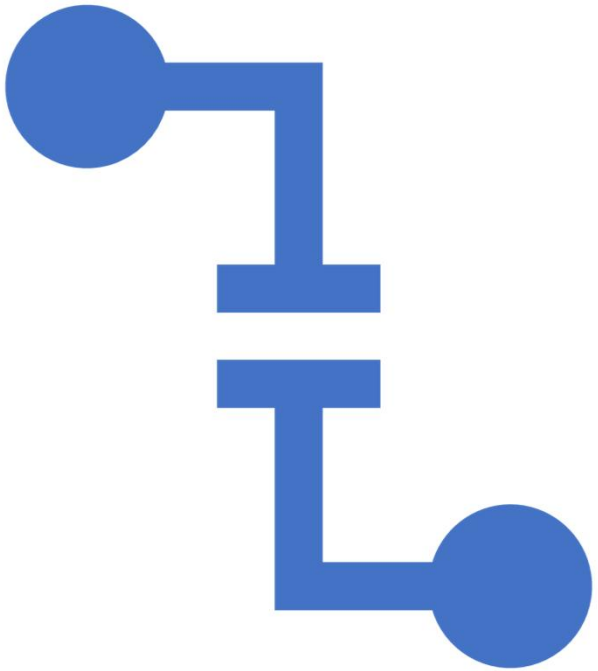
PROF. FERNANDO AMARAL
www.datascientist.com.br

Elementos

- SparkSession: Seção
- Aplication: 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