

# O Que é Spark

- Ferramenta de Processamento de Dados (Não é Data Storage)
- Distribuído em um Cluster
- Em memória
- Veloz
- Escalável
- Particionamento

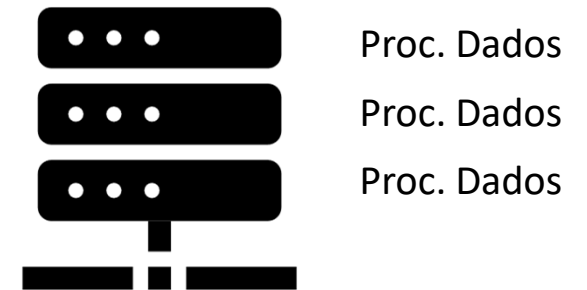


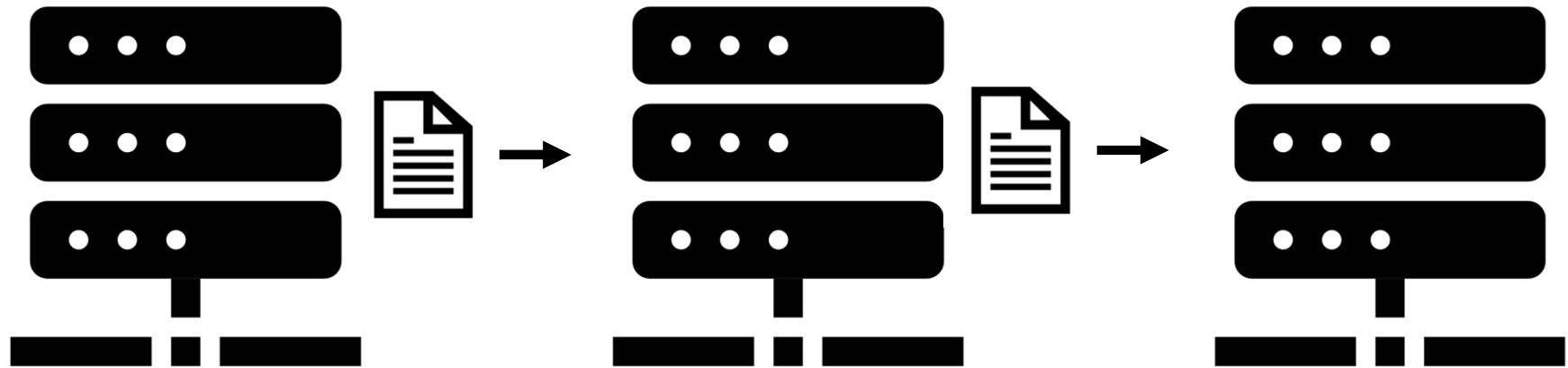
# Cluster

- Rede de Computadores



- Rede de Computadores - Cluster

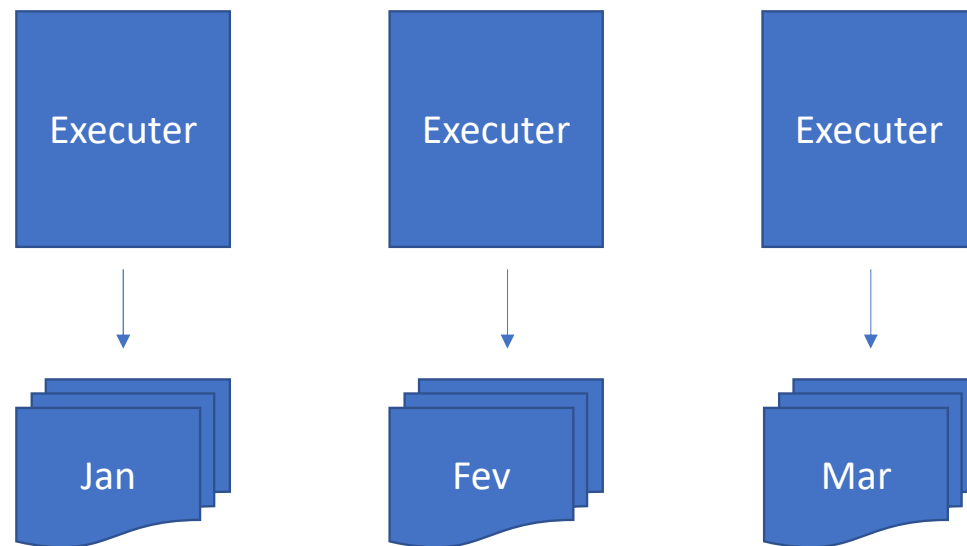




## Replicação e Tolerância a Falha

- 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 

---

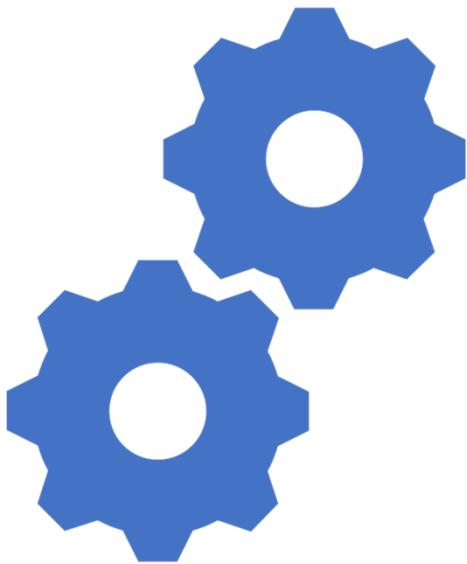
R 

---

SQL 

# Projeto “Extremamente” Ativo

+ 1000 colaboradores ativos



## Spark

Universidade da Califórnia iniciou projeto Spark em 2009

Versão 1.0 lançada em Maio de 2014 pela Fundação Apache