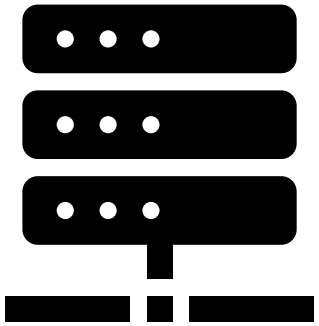
Introdução ao Spark

O Que é Spark

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

Cluster

- Rede de Computadores

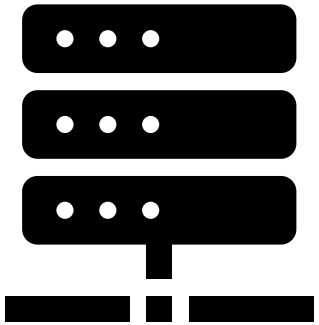


E-mail

Storage

ERP

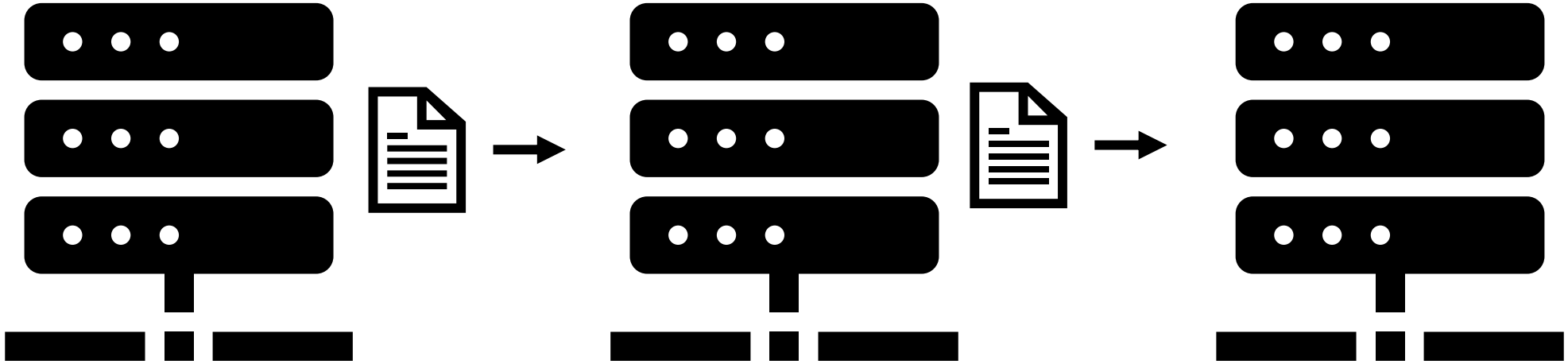
- Rede de Computadores - Cluster



Proc. Dados

Proc. Dados

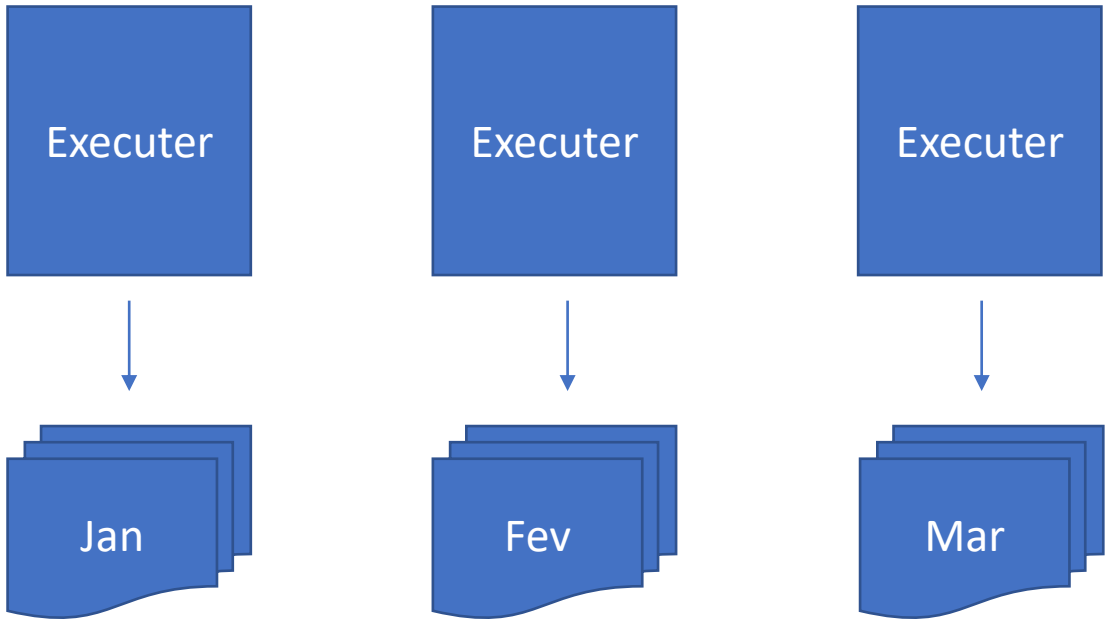
Proc. Dados



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

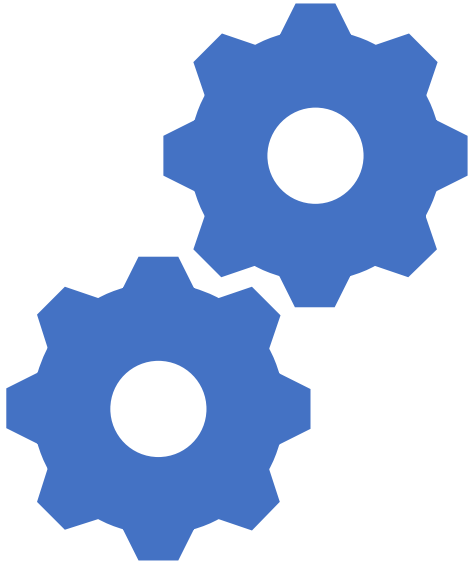
Scala 

Python 

Java 

R 

SQL 



- + 1000 colaboradores ativos

Projeto “Extremamente” Ativo

História

- Google File System (GFS), MapReduce (MR) e Bigtable

- Resultou em Hadoop, MapReduce, HDFS e Yarn
  - Complexo
  - Requer conhecimento em Java
  - Modelo em Batch em tarefas Mapeamento e Redução

- Solução
  - Ex: Hive criado pelo Facebook: DW SQL sobre HDFS

Spark

Universidade da Califórnia iniciou projeto Spark em 2009

Versão 1.0 lançada em Maio de 2014 pela Fundação Apache