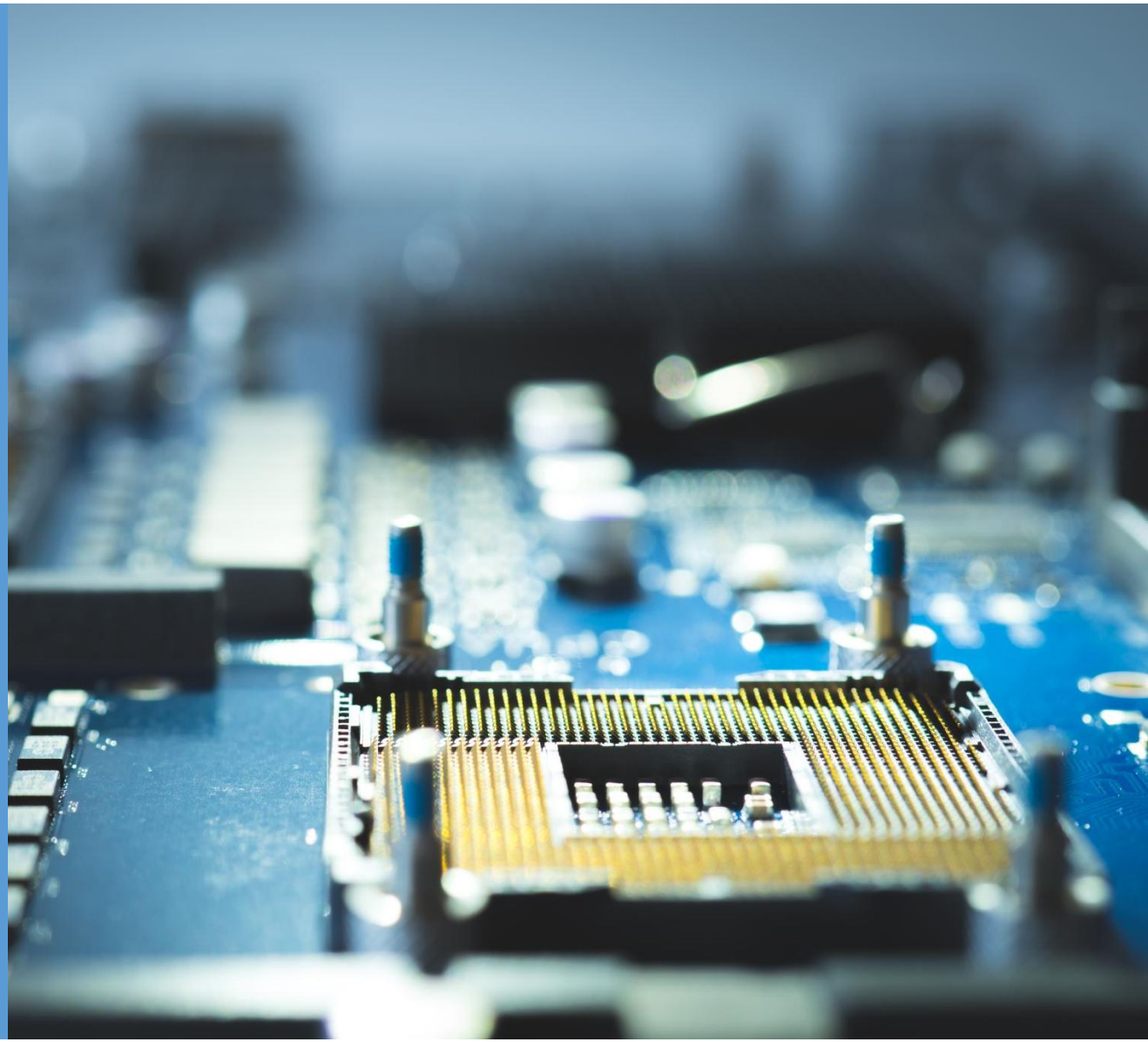
O que é Spark?

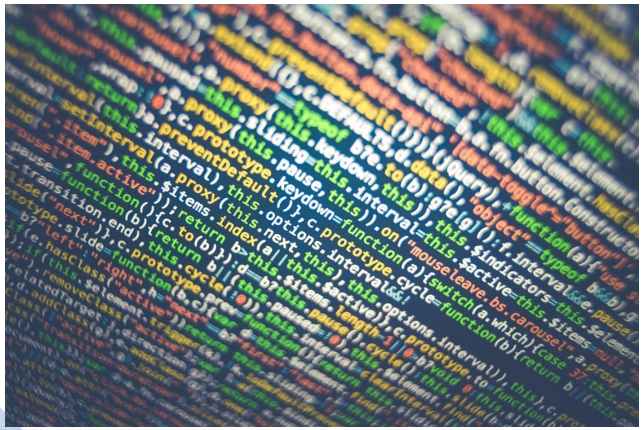
Dados!

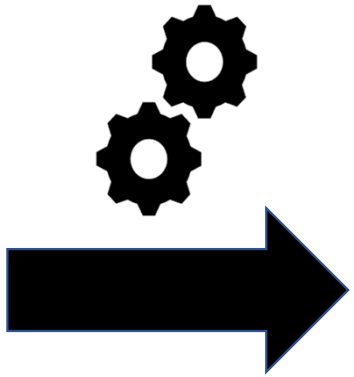
- Armazenamento
- Processamento



Processamento?

- Produzir valor!

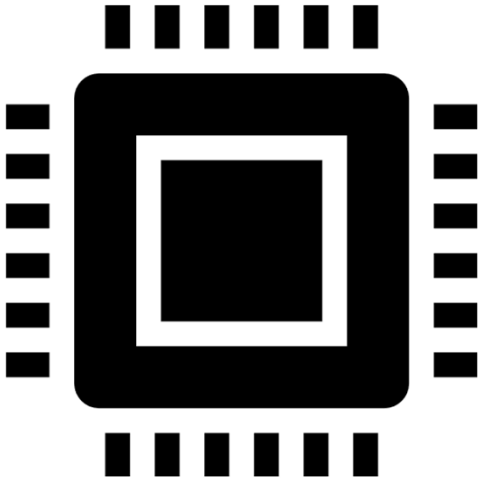




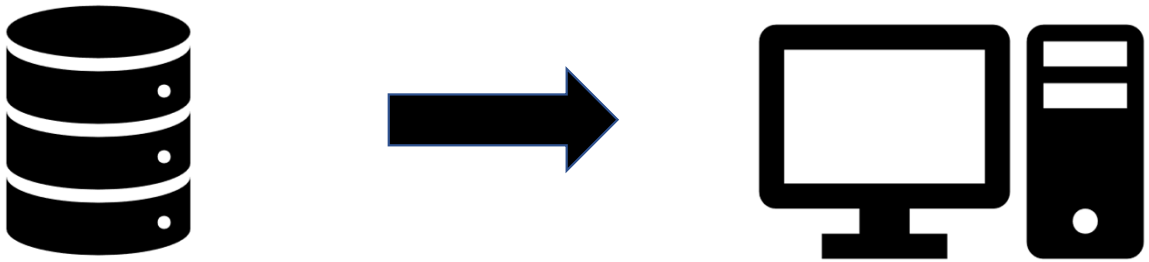
Processar Dados

- Poder Computacional!

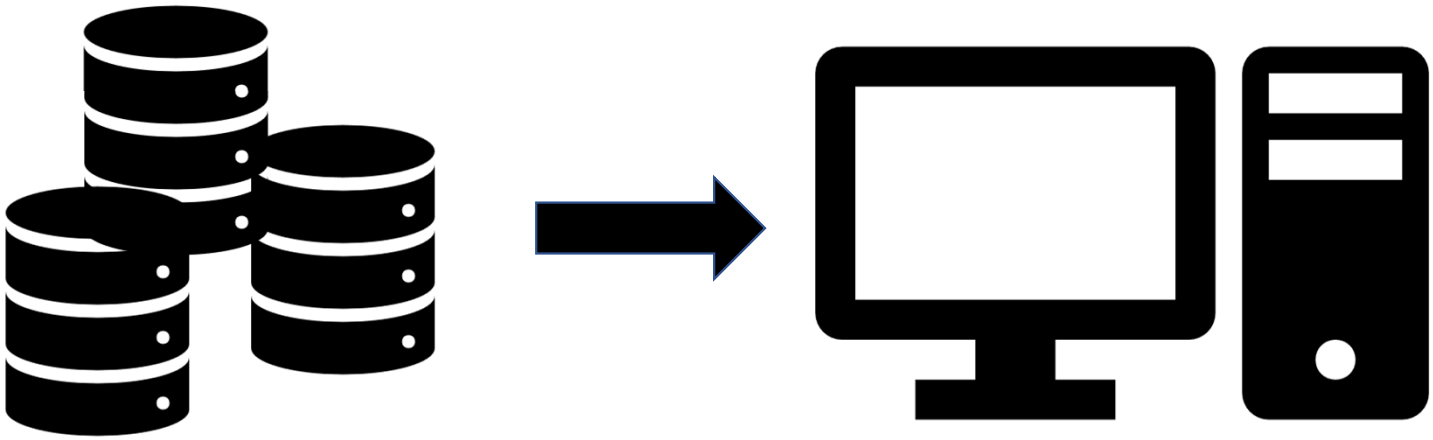
- CPU
- Memória
- Disco



Dados



Mais Dados!



Big Data!

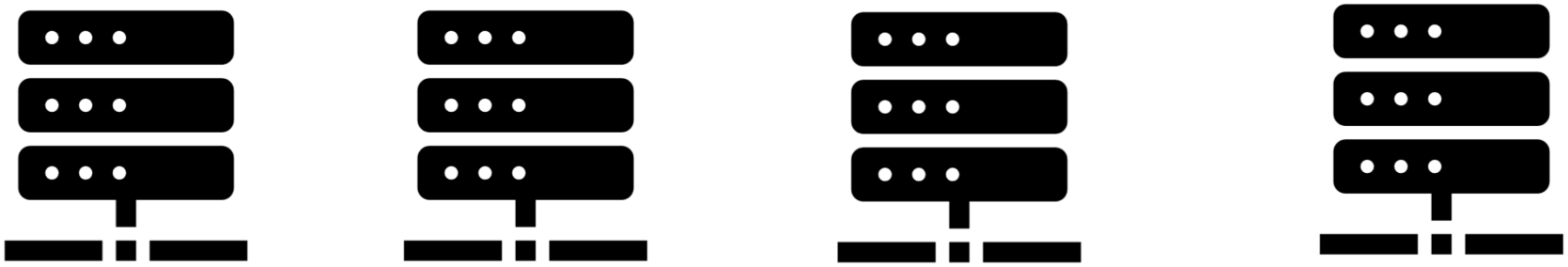
Existe um limite!

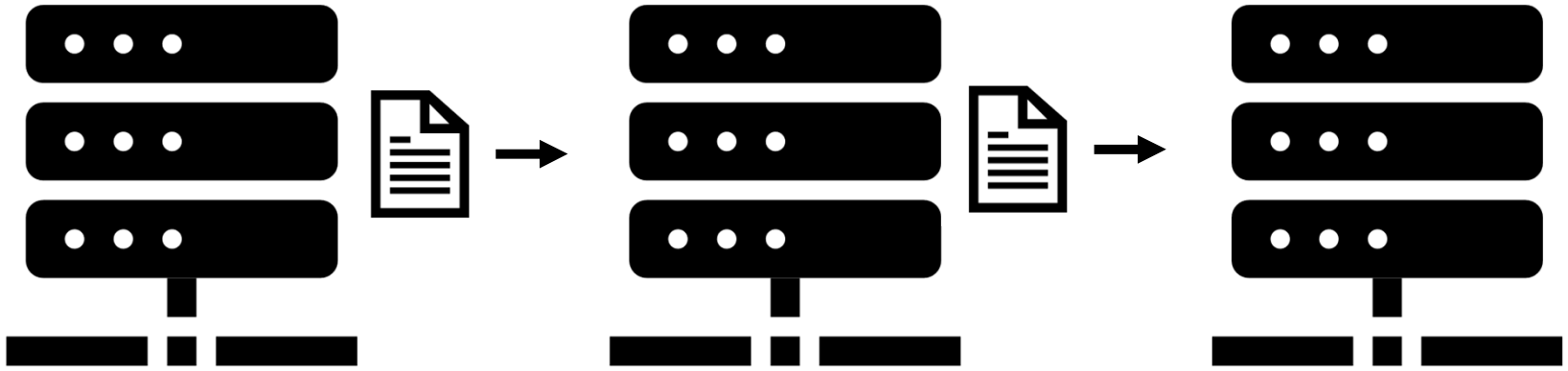
Spark

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

Spark

- Escala horizontalmente - Cluster

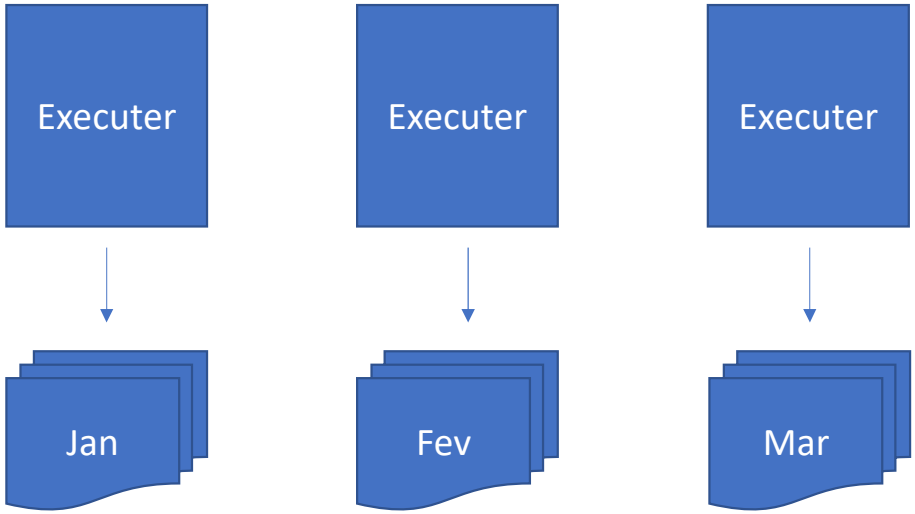




Replicação / Tolerância a Falha

- 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?

- Aprendemos a processar dados, criar modelos etc. com Python e R, utilizando bibliotecas como Pandas, Scikit Learn etc.
- Precisamos de Spark?
  - Alta performance pela sua natureza “distribuída”
- Com Pyspark, você tem tudo do Python + Spark!