



# Spark

---

Memória

Operação em Cluster

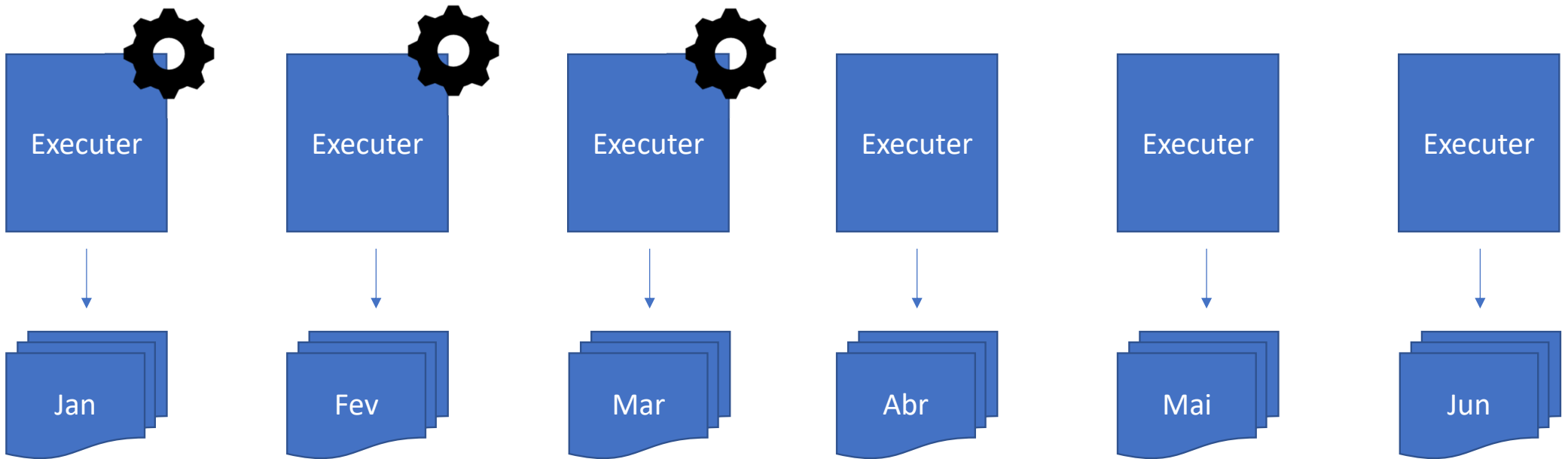
Particionamento (divisão de dados)

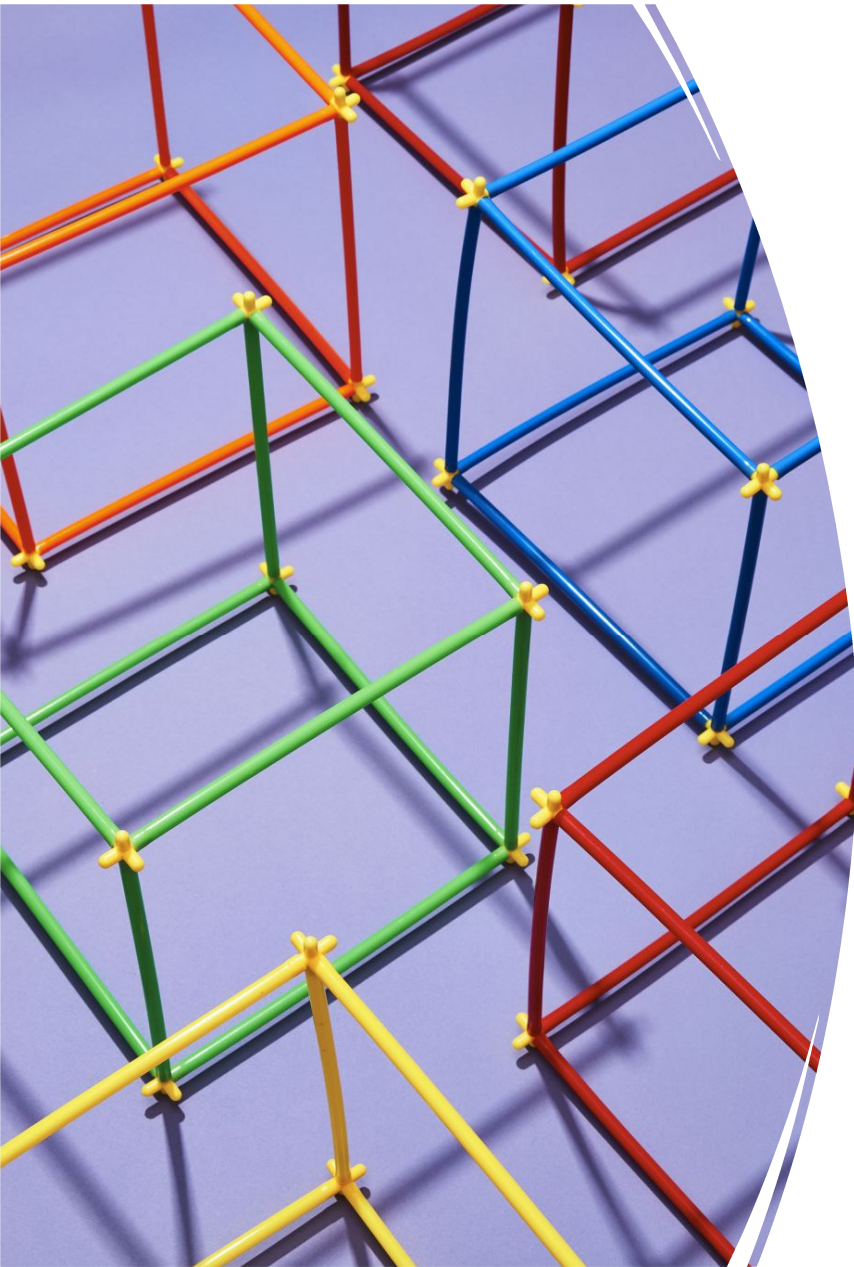
Paralelismo

Redundância

# Particionamento

## Total de vendas de Janeiro a Março





# Particionamento

---

Por padrão dados são particionados de acordo com número de núcleos

Cada partição fica em um nó e tem uma task



# Shuffle

---

Redistribuição de Dados entre partições

# Particionamento

---

- Dados são particionados por padrão e dependem de vários fatores e configurações
- Podemos particionar explicitamente em disco (`partitionBy`)
- Ou em memória `repartition()` or `coalesce()`

# Bucketing

Semelhante a particionamento, porém com número fixo de partições

Ideal para coluna com alta cardinalidade

Pode ser usado com conjunto com Particionamento