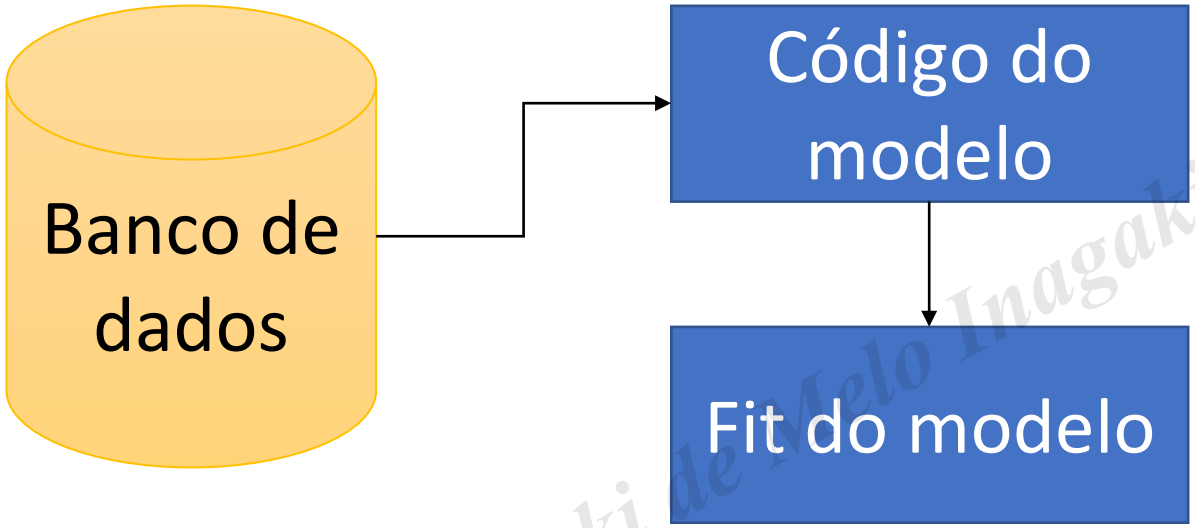
MBA
USP
ESALQ

BIG DATA E DEPLOYMENT DE MODELOS

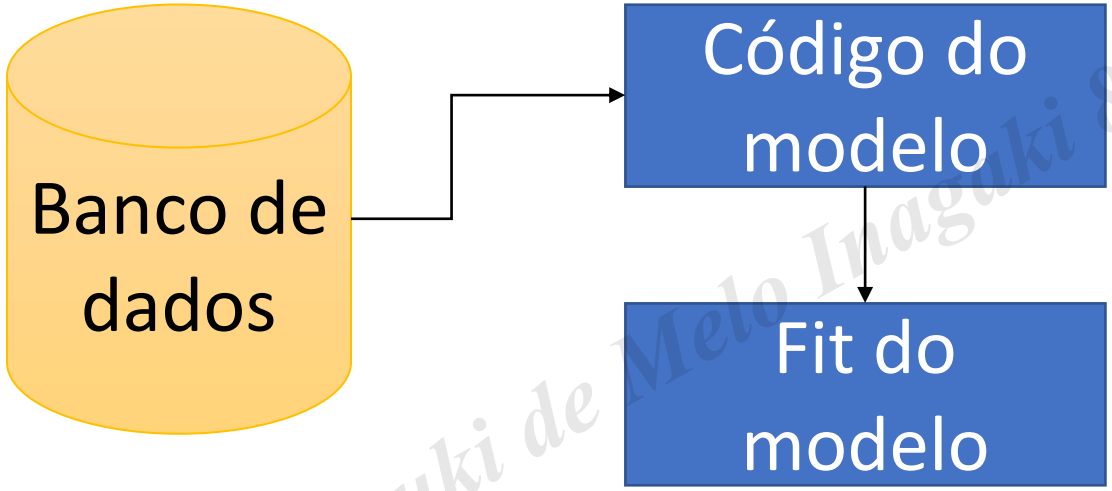
Helder Prado Santos

*A responsabilidade pela idoneidade, originalidade e licitude dos conteúdos didáticos apresentados, é do professor.

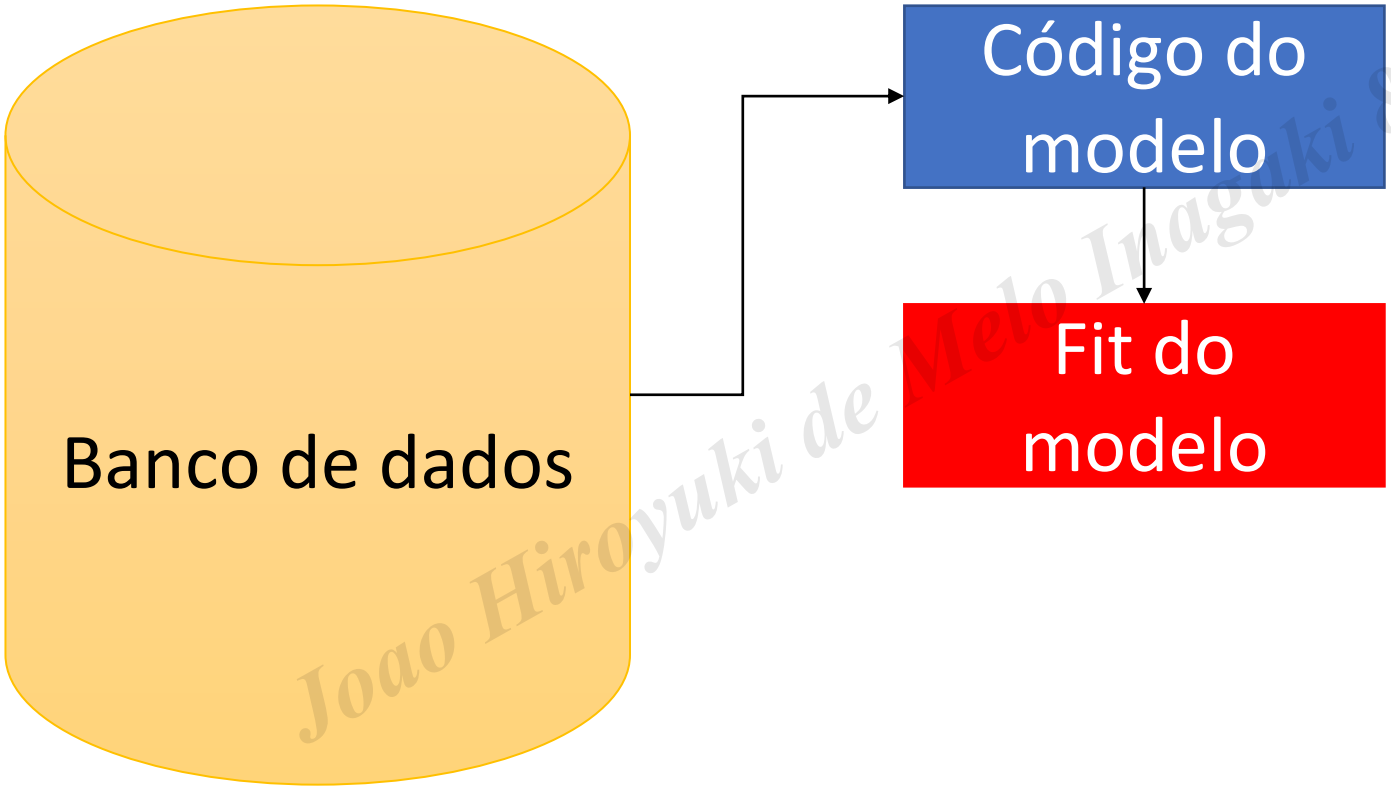
Vamos desenvolver um modelo



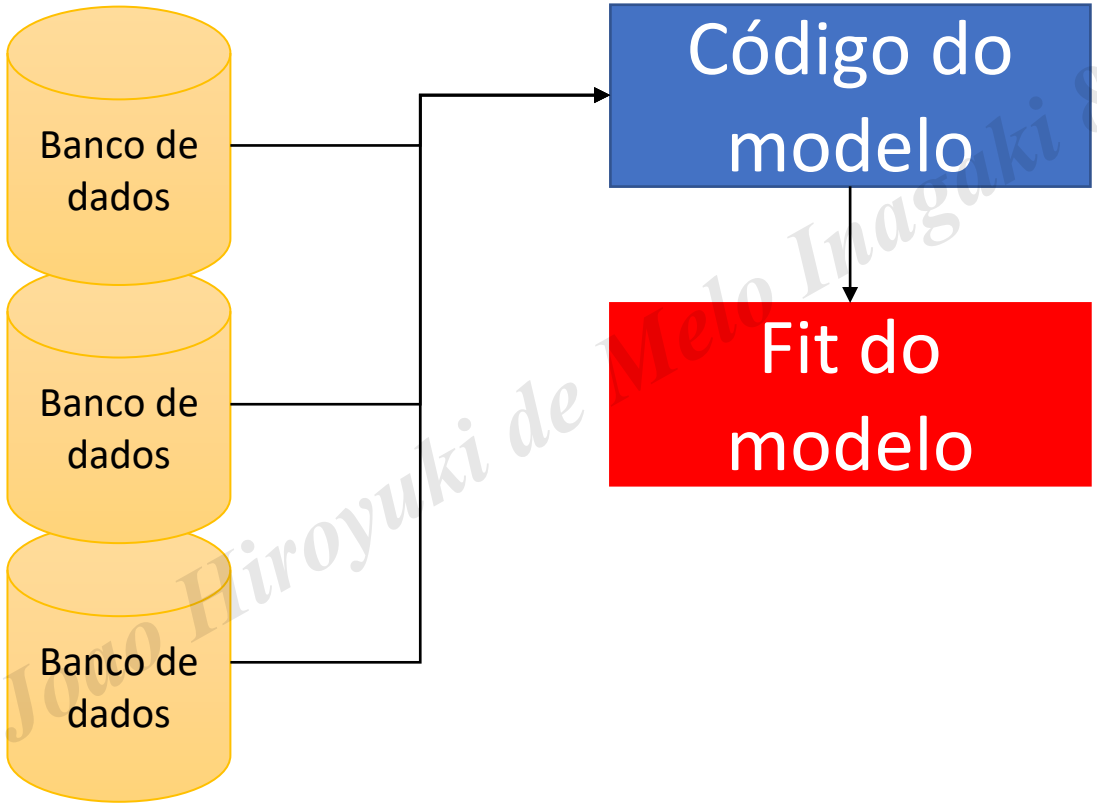
E se?



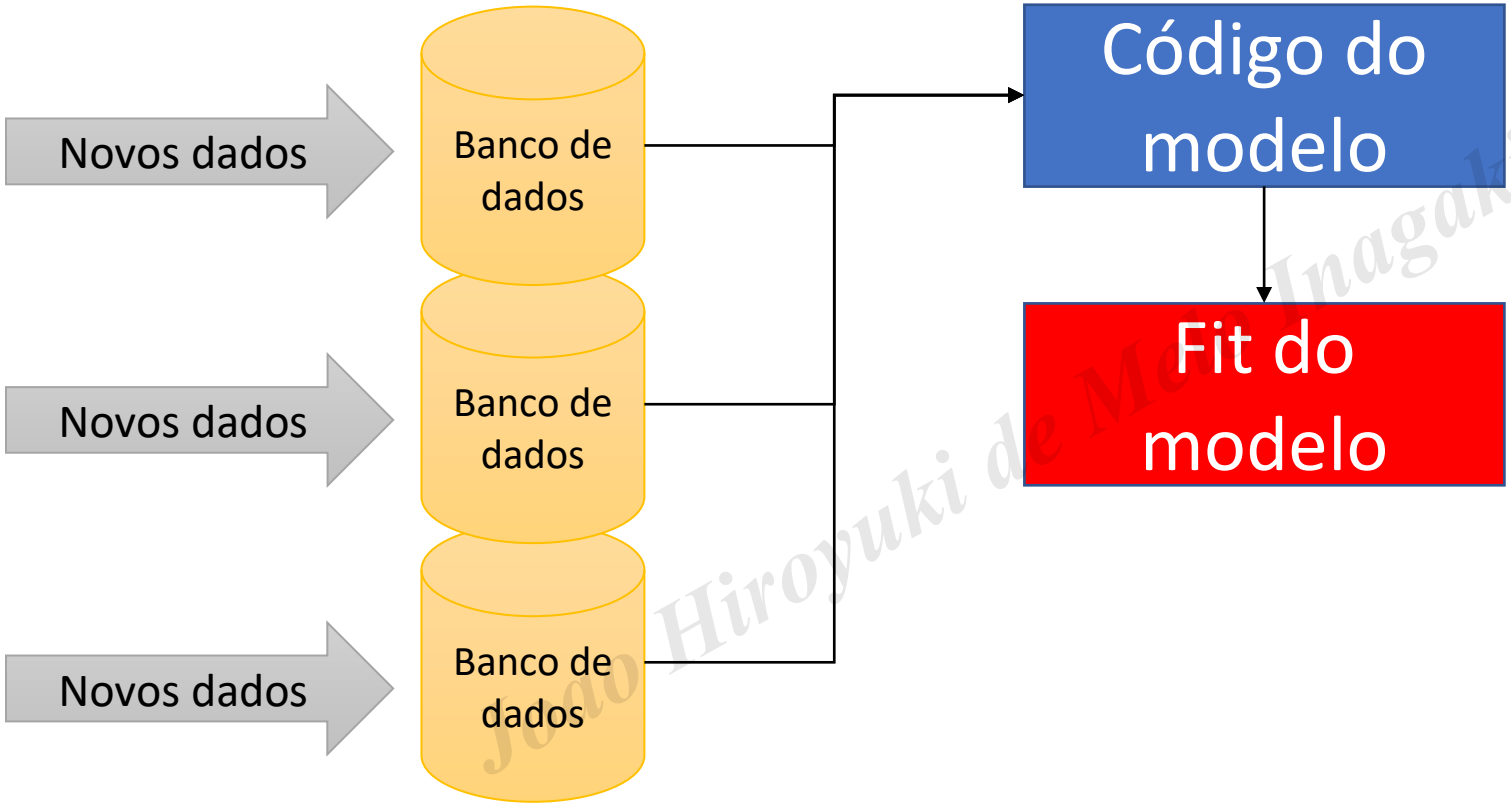
E se?



E se?



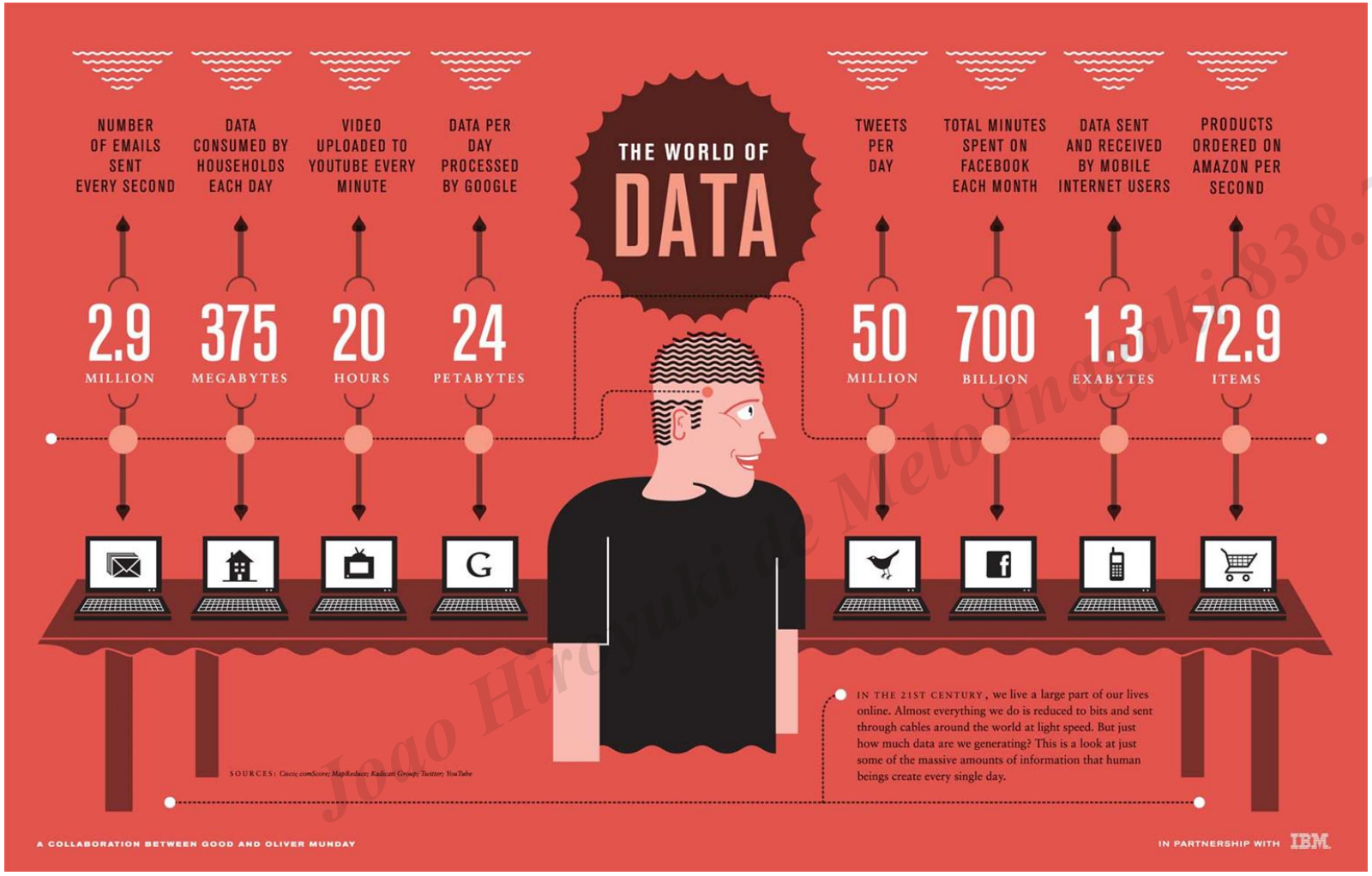
E se?



Alguns problemas encontrados...

1. E se nosso banco de dados crescer?
2. E se tivermos vários bancos de dados?
3. E se novos dados chegarem ao longo do tempo?

O nosso mundo de dados



Fonte: https://hexanika.com/history-of-big-data/

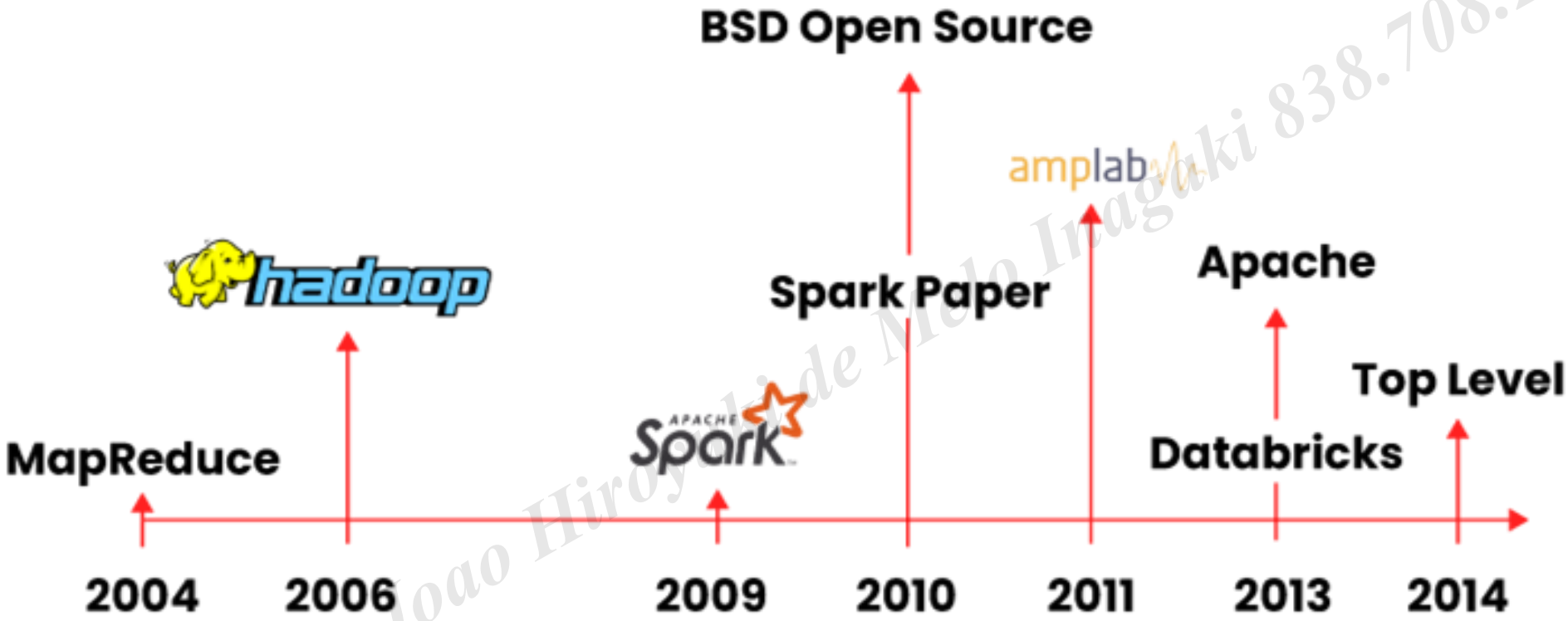
Entrando no mundo do Big Data

- É uma ferramenta?
- É só ter grandes volumes de dados?

Big Data: Os 5V

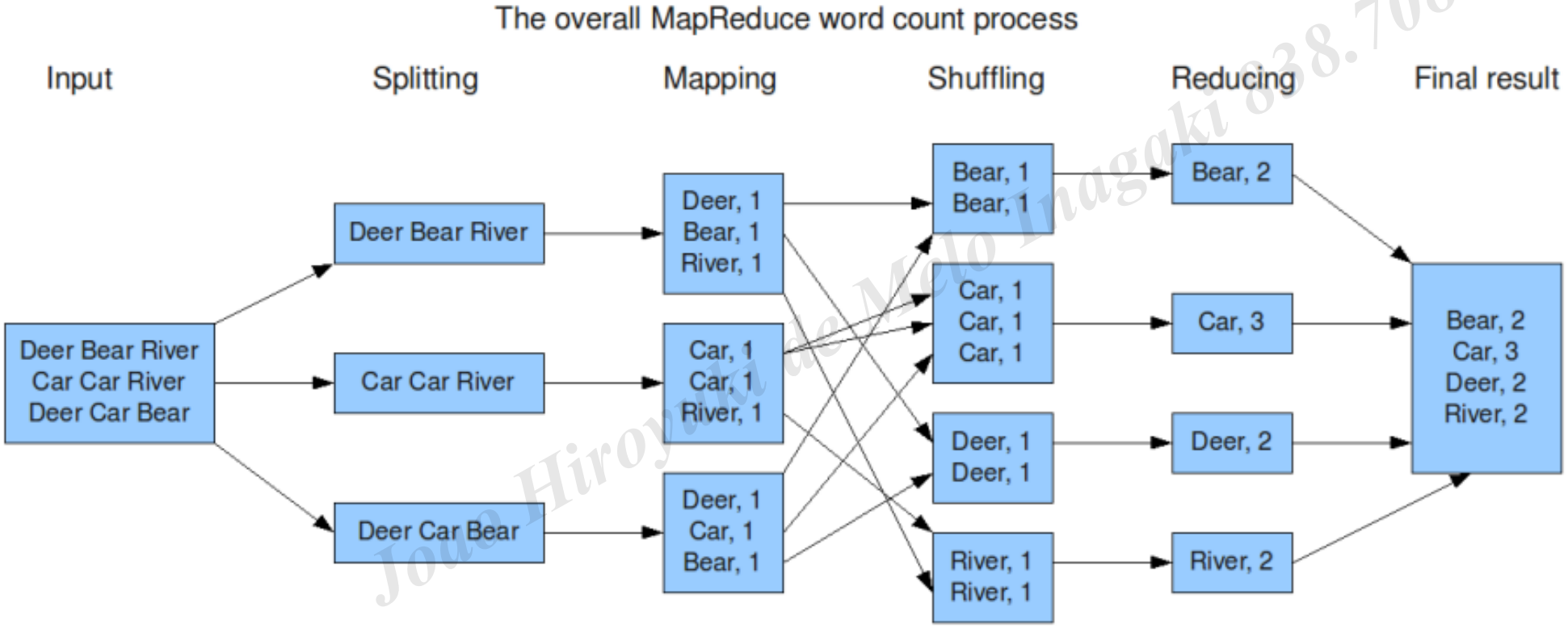
Fonte: https://www.cortex-intelligence.com/blog/os-5-vs-do-big-data

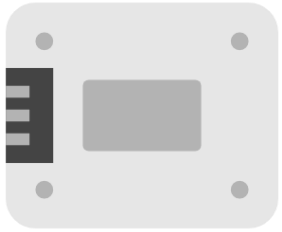
Sua história e evolução



Fonte: https://betravingknows.com/weekly-reports/data-analytics/2018/05/how-big-data-shaped-todays-casino-and-why-you-should-care/

Hadoop e a introdução do MapReduce



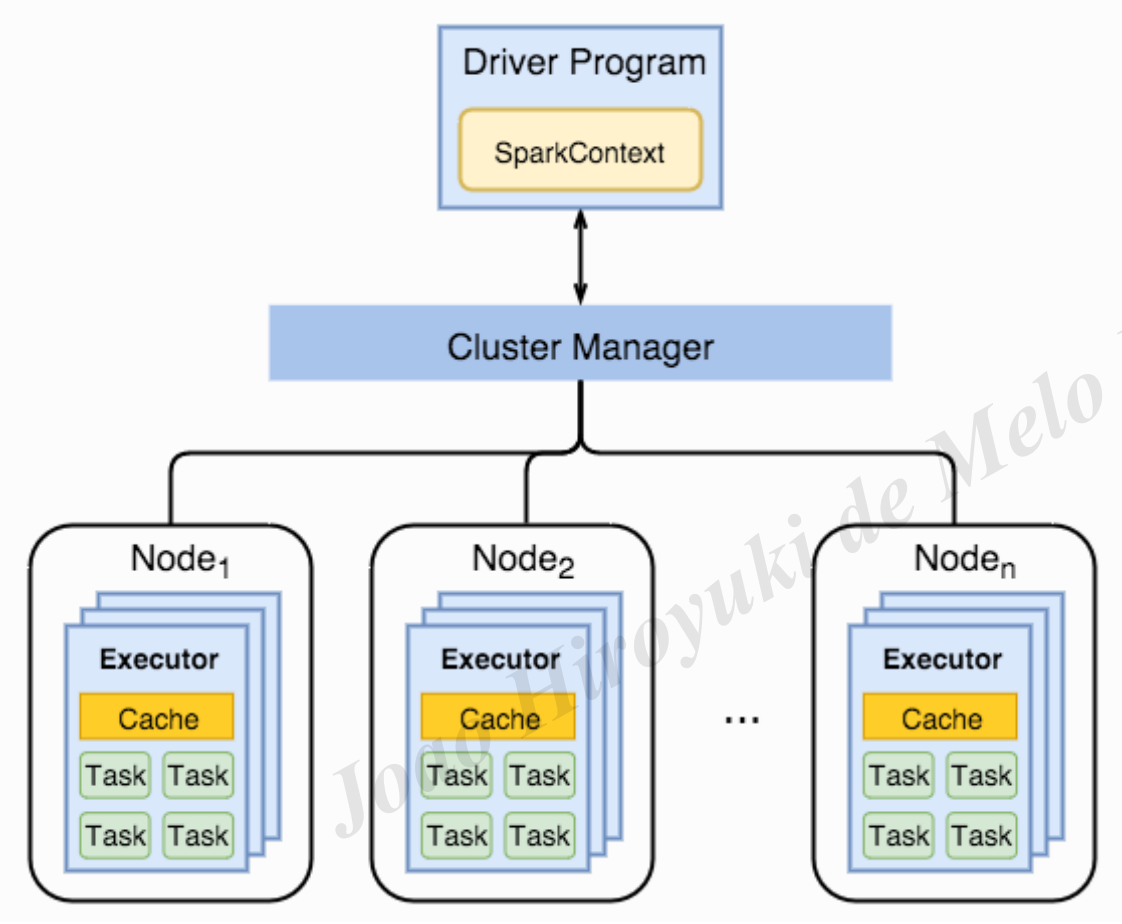


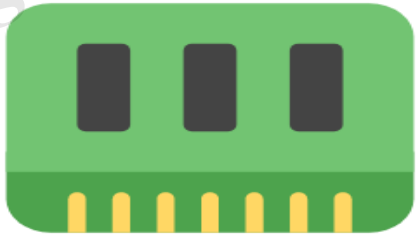
Processamento em disco

Fonte: https://betravingknows.com/weekly-reports/data-analytics/2018/05/how-big-data-shaped-todays-casino-and-why-you-should-care/

Spark e a introdução do RDD

Resilient distributed datasets

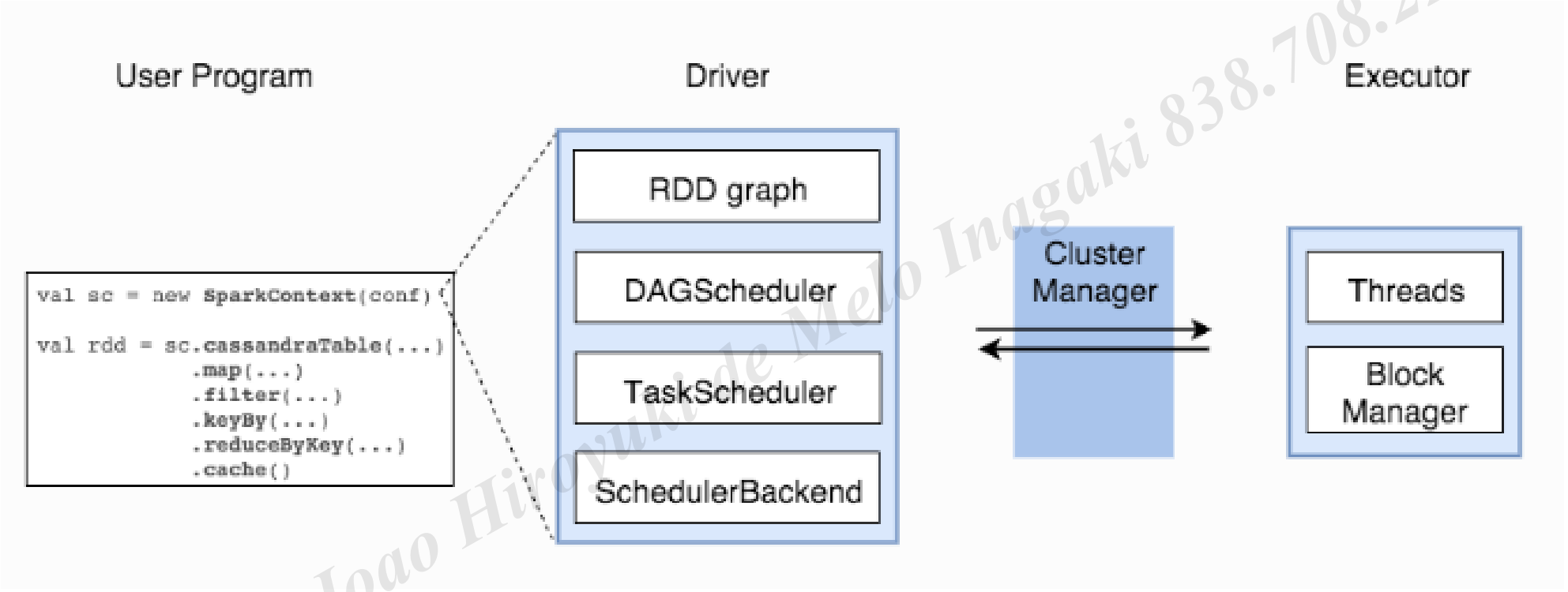




Processamento em memória

Fonte: https://runawayhorse001.github.io/LearningApacheSpark/introduction.html

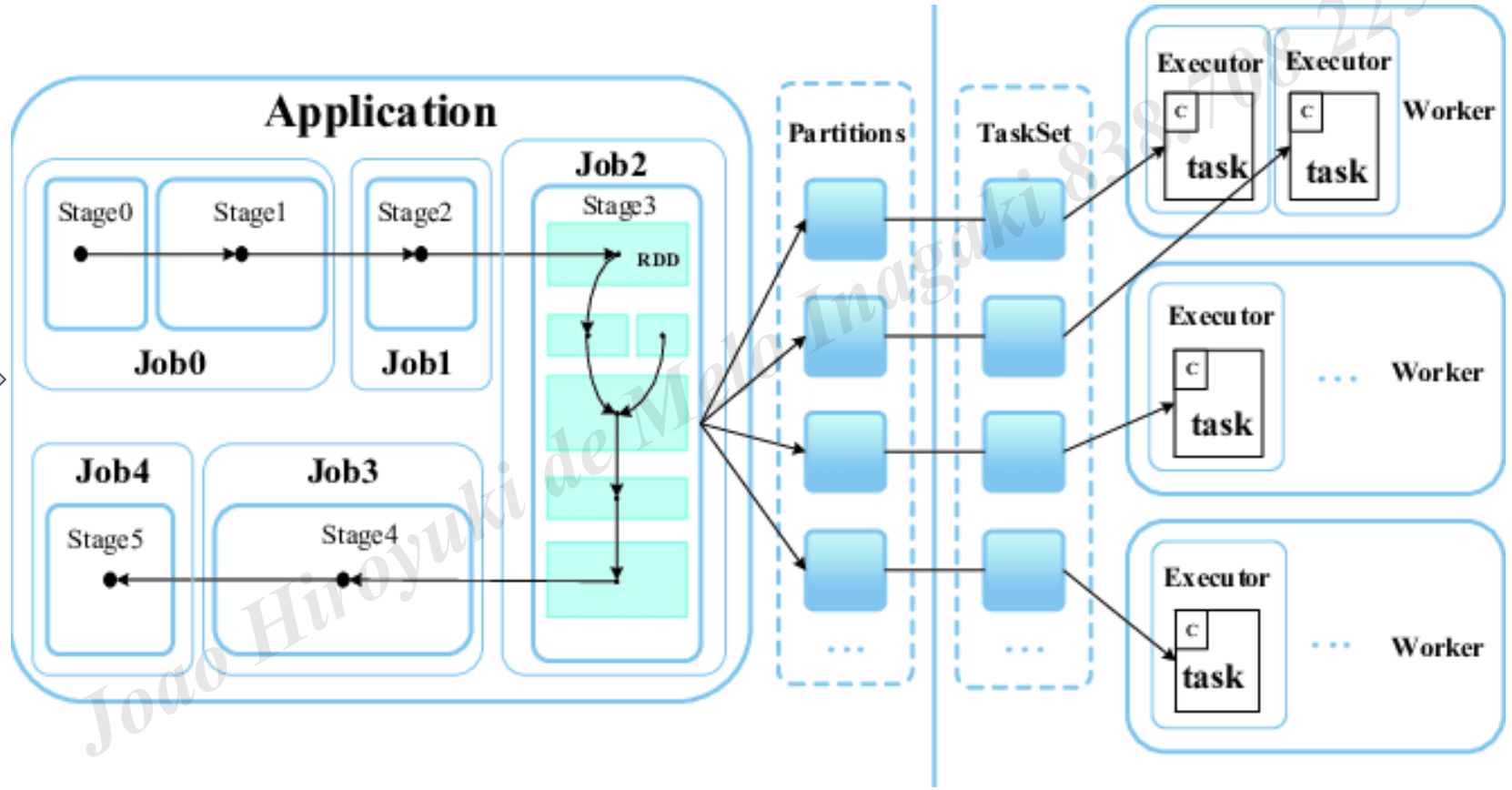
Spark e a introdução do RDD



Fonte: https://runawayhorse001.github.io/LearningApacheSpark/introduction.html

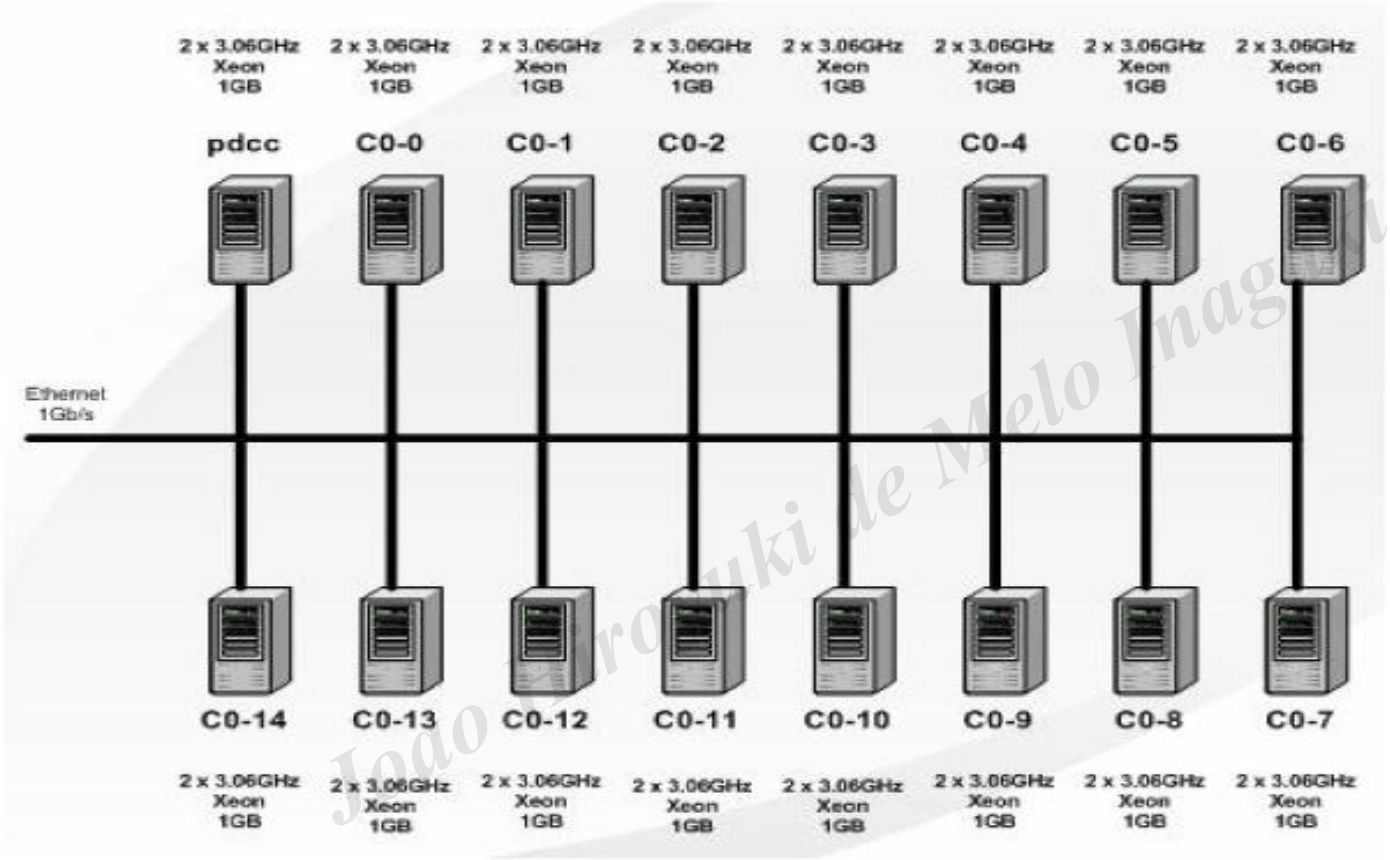
Spark e a introdução do RDD

```
rdd1.join(rdd2)
    .groupBy(…)
    .filter(…)
```



Fonte: https://runawayhorse001.github.io/LearningApacheSpark/introduction.html

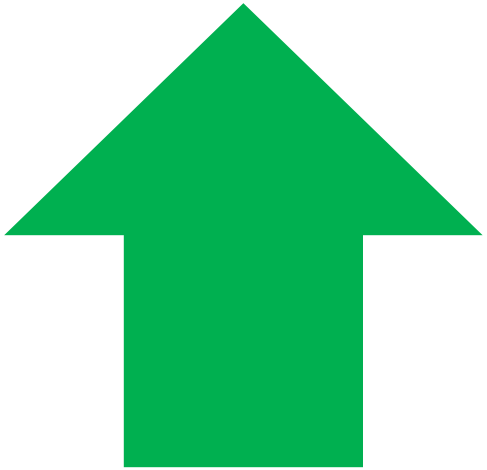
Possibilidade de criar cluster de processamento



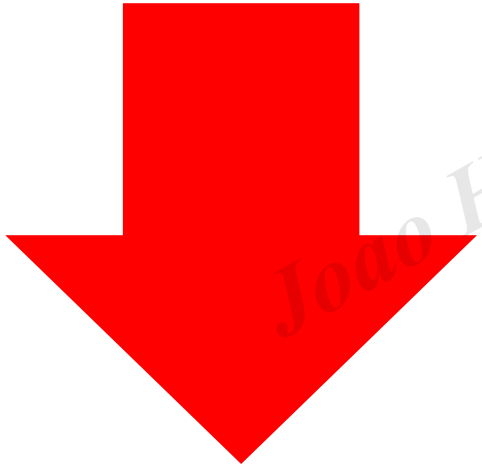
Poder de processamento

Fonte: https://runawayhorse001.github.io/LearningApacheSpark/introduction.html

Estratégias e boas práticas



Performance



Custos

1. Usar um framework de big data
2. Comprimir os dados
3. Particionar os dados

Pensar na compressão de dados

Parquet

... Até **87%** de redução de tamanho

... Até **34x** mais rápido para carregar os dados

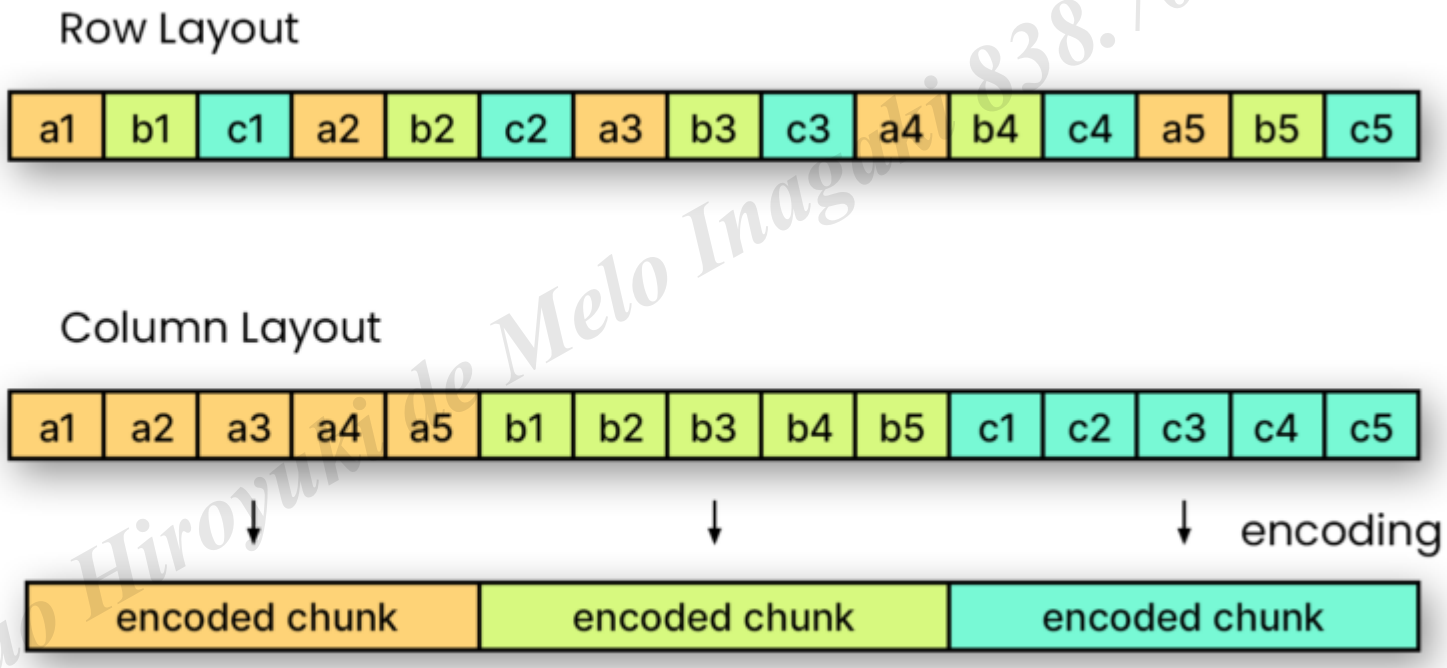
... Até **99%** de redução de custos

Fonte: https://www.databricks.com/glossary/what-is-parquet

Como ele funciona?

Logical table representation

a	b	c
a1	b1	c1
a2	b2	c2
a3	b3	c3
a4	b4	c4
a5	b5	c5



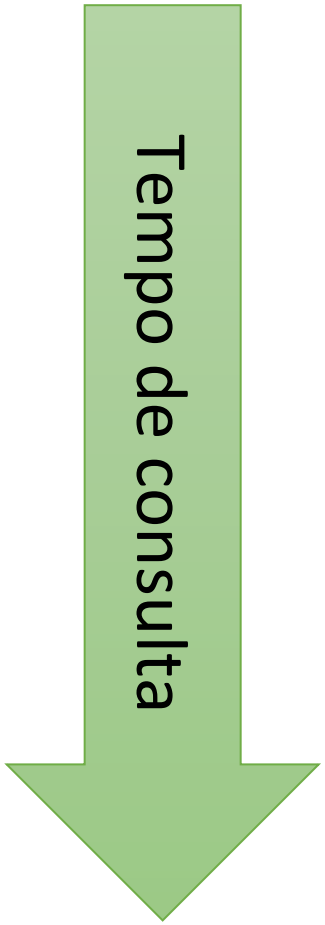
Fonte: https://www.dremio.com/resources/guides/intro-apache-parquet/

Comprime mesmo?

Dataset	Size on Amazon S3	Query Run time	Data Scanned	Cost
Data stored as CSV files	1 TB	236 seconds	1.15 TB	$5.75
Data stored in Apache Parquet format*	130 GB	6.78 seconds	2.51 GB	$0.01
Savings / Speedup	87% less with Parquet	34x faster	99% less data scanned	99.7% savings

Fonte: https://blog.openbridge.com/how-to-be-a-hero-with-powerful-parquet-google-and-amazon-f2ae0f35ee04

Particionar os dados



.../tabela_de_vendas/

*.../tabela_de_vendas/ano=**2022***

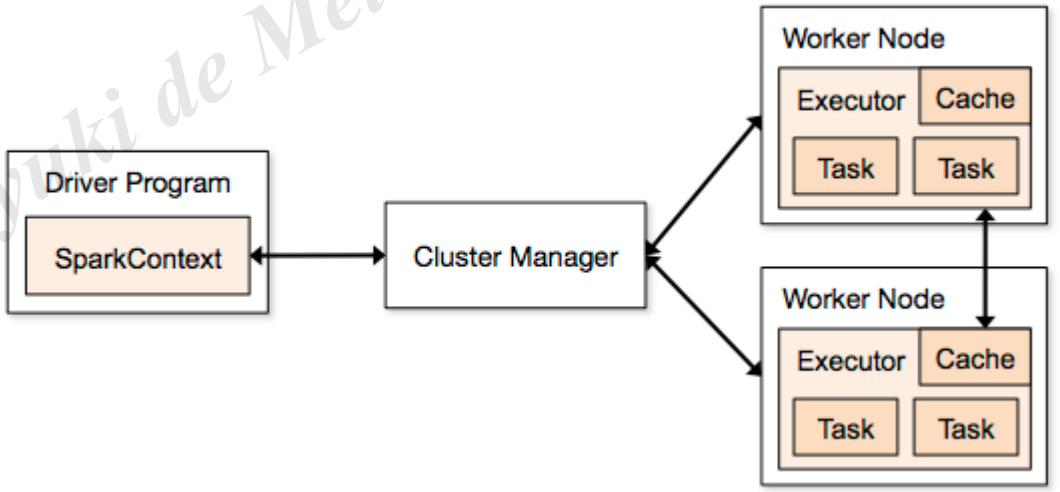
*.../tabela_de_vendas/ano=**2022**/mes=**8***

Drasticamente!!!!

... E em 15 anos?

Framework que iremos utilizar

- Spark
  - Processamento distribuído em memória
  - Ferramenta open source
  - Possui biblioteca built-in de Machine Learning



Fonte: https://spark.apache.org/docs/1.1.0/cluster-overview.html

Mãos à obra!

OBRIGADO!

linkedin.com/in/helderprado