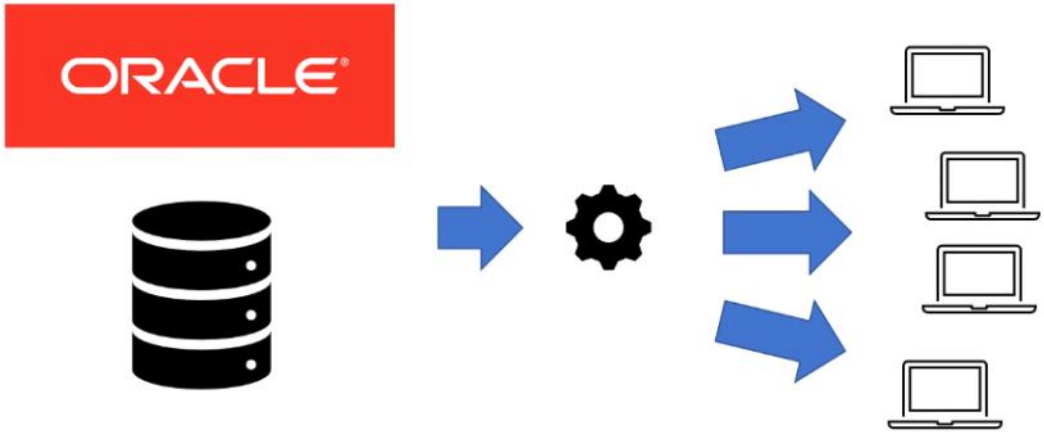
04.Formatos de Big Data

quinta-feira, 21 de setembro de 2023 15:15

Armazéns de dados clássicos

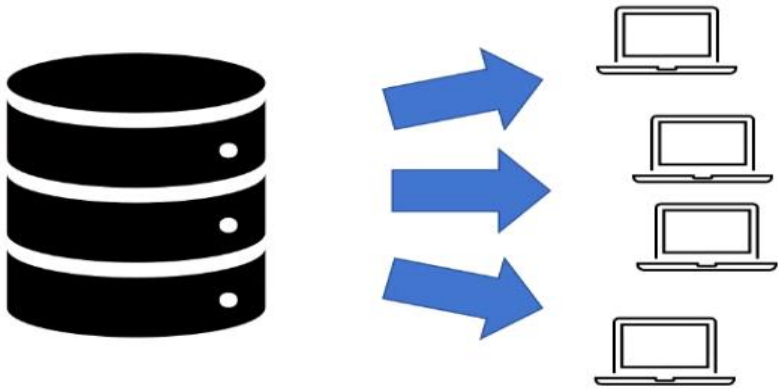
- **Formatos antigos (Proprietários)**



- **Precisava de driver**

Armazéns de dados modernos

- **Formatos abertos**



- **Qualquer ferramenta pode acessar os dados**

- Armazéns de dados modernos armazenam os dados em formatos desacoplados de ferramentas e abertos.
- Formatos binários, compactados
- Suportam Schema
- Podem ser particionados entre discos:
  - Redundância
  - Paralelismo

Formatos de dados

- **Parquet** - Colunar, padrão do spark
- **ORC** - Colunar, padrão do Hive
- **Avro** - Linha

- Muitos atributos e mais escrita - Linha
- Menos atributos e mais leituras - Coluna

- Em geral ORC é mais eficiente na criação(escrita) e na compreensão
- Parquet tem melhor performance na consulta(leitura)
- O ideal é fazer um benchmark