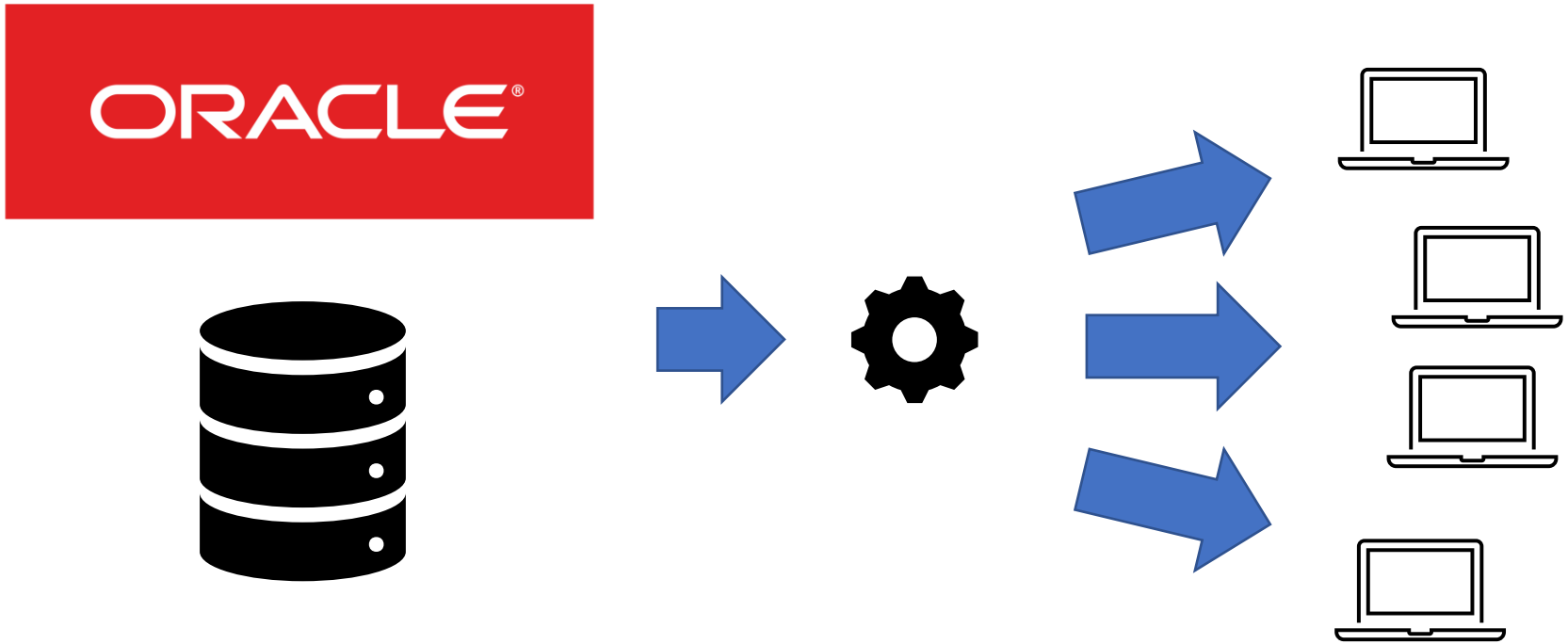
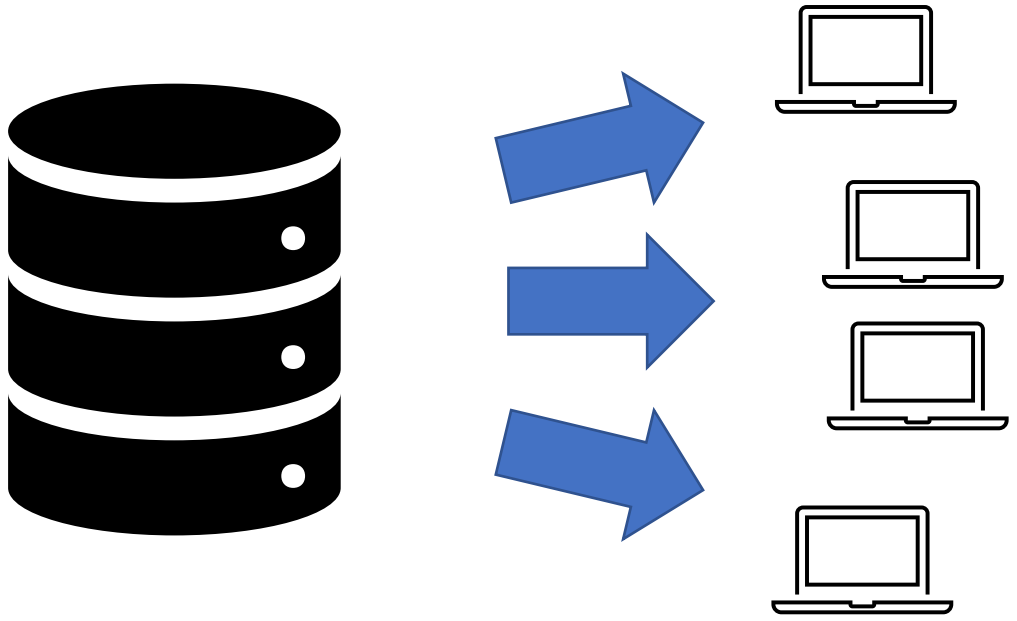
Formatos de Big Data

Armazéns de Dados Clássicos
Formatos Proprietários



Armazéns de Dados Modernos
Formatos Abertos



Formatos para Big Data

Apache orc™

PROF. FERNANDO AMARAL
www.datascientist.com.br

APACHE Spark™

Formatos para Big Data

- Armazéns de dados modernos tendem a armazenar dados em formatos “desacoplados” de ferramentas e abertos
- Formatos binários, compactados
- Suportam Schema
- Podem ser particionados entre discos:
  - Redundância
  - Paralelismo

Formatos

- Parquet – Colunar, padrão do Spark
- ORC – Colunar, padrão do Hive
- Avro – Linha

- Muito atributos e mais escrita – linha
- Menos atributos e mais leitura, coluna

Qual escolher?

- Em geral ORC é mais eficiente na criação (escrita) e na compressão
- Parquet tem melhor performance na consulta (leitura)

- O ideal é fazer um benchmark!