



Formatos para Big Data



Formatos para Big Data

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



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!

