



Ficou curioso?

A quantidade de dados que se usa, quando chega ao nível de *big data*, pode ser bem complicada. As soluções conhecidas podem não mais funcionar.

Por exemplo, é comum armazenar dados em bancos de dados relacionais, porém, à medida que se vai avançando pra casa de *Tera bytes*, o banco não vai ter a mesma *performance* de antes? Nisso o processo para criar índices e otimizá-lo tem de ser feito com mais cautela, ou pagar o preço por uma máquina maior. Talvez o melhor seja estudar a forma como esses dados sejam consumidos e tentar outras formas.

Existem outras formas de armazenar esses dados, como por exemplo o ORC e o PARQUET, mas note que não existe almoço grátis. A forma como são operam e são construídas podem torná-las vantajosas por um lado e ruins por outro. Sempre vale a pena pesquisar bastante e fazer provas de conceito antes de criar o seu sistema.

A AWS tem uma ferramenta para esse tipo de trabalho em big data que trabalha inclusive com ferramentas de código aberto, seu nome é AWS EMR.