



Big Data

INSIGHTS · DATA ENGINEERING

**Um Lakehouse mal
particionado custa
mais em consultas do
que toda a ingestão.**

Desliza →

O problema em 1 frase

Sem partição adequada, queries podem ler 5–10 TB quando só precisam do último mês, multiplicando custos e latências.

Por que particionar?

Partition pruning faz o sistema ler apenas as pastas relevantes (ex.: year/month), reduzindo I/O e tempo de resposta.

Como escolher a chave

Priorize colunas usadas em filtros frequentes (mês, país). Evite partições com alta cardinalidade (>10k–100k valores).

Granularidade certa

Objetivo prático: ficheiros médios entre 128–512 MB. Se o médio ficar abaixo de 50 MB, é sinal para compactar.

Quando compactar

Após cargas que geram milhares de ficheiros (ex.: 1 TB em 20k ficheiros) ou quando uma partição tem >5k–10k ficheiros.

Estratégias de compactação

Minor compaction junta pequenos ficheiros;
major compaction reescreve e aplica
ordenação/Z-order para queries
multidimensionais.

Implementação no Fabric

Defina partições, ingerir em staging, agende notebooks Spark para compactação e execute VACUUM conforme política de retenção.

CONTINUA NO BLOG

Por que particionar um Lakehouse?

www.bconcepts.pt/insights