

A ingestão não é mover dados — é garantir que cada byte novo conta.

Desliza →

Porquê ingestão incremental

Reduz latência, custos e risco de inconsistências. Uma carga completa de 1 TB pode custar €800–€1.500 por execução; o incremental processa tipicamente 20–50 GB por dia.

Princípios essenciais

Idempotência, rastreabilidade e limites de janela.
Use chaves naturais, timestamps e batch_id para
garantir resultados previsíveis.

CDC, watermarking e checkpoints

CDC é o padrão quando disponível. Sem CDC, aplique watermark por partição e tolerâncias de lateness (1–5 minutos ou mais, conforme SLA).

Tabela de controle típica

Registe batch_id, source_system, start_time, end_time, record_count, bytes, status, error_message e processed_at para auditoria e reproprocessamento selectivo.

Deduplicação e MERGE

MERGE garante atomicidade mas é custoso em tabelas enormes. Hash de linha (SHA-256) pode reduzir IO 40–70% quando poucas linhas mudam.

Arquitetura recomendada

Conectores → zona raw no Lakehouse → job Spark para merge idempotente → actualização de checkpoints → tabelas otimizadas para consumo.

Mini-caso: retalhista nacional

80 lojas: cargas nocturnas de 1 TB causavam 4h de refresh e ~€1.200. Com CDC e checkpoints o delta ficou ~30 GB/dia e a latência caiu para minutos.

CONTINUA NO BLOG

Como implementar ingestão incremental resiliente no Microsoft Fabric?

www.bconcepts.pt/insights