5 Methods to Optimize Lakehouse

01 Partitioning

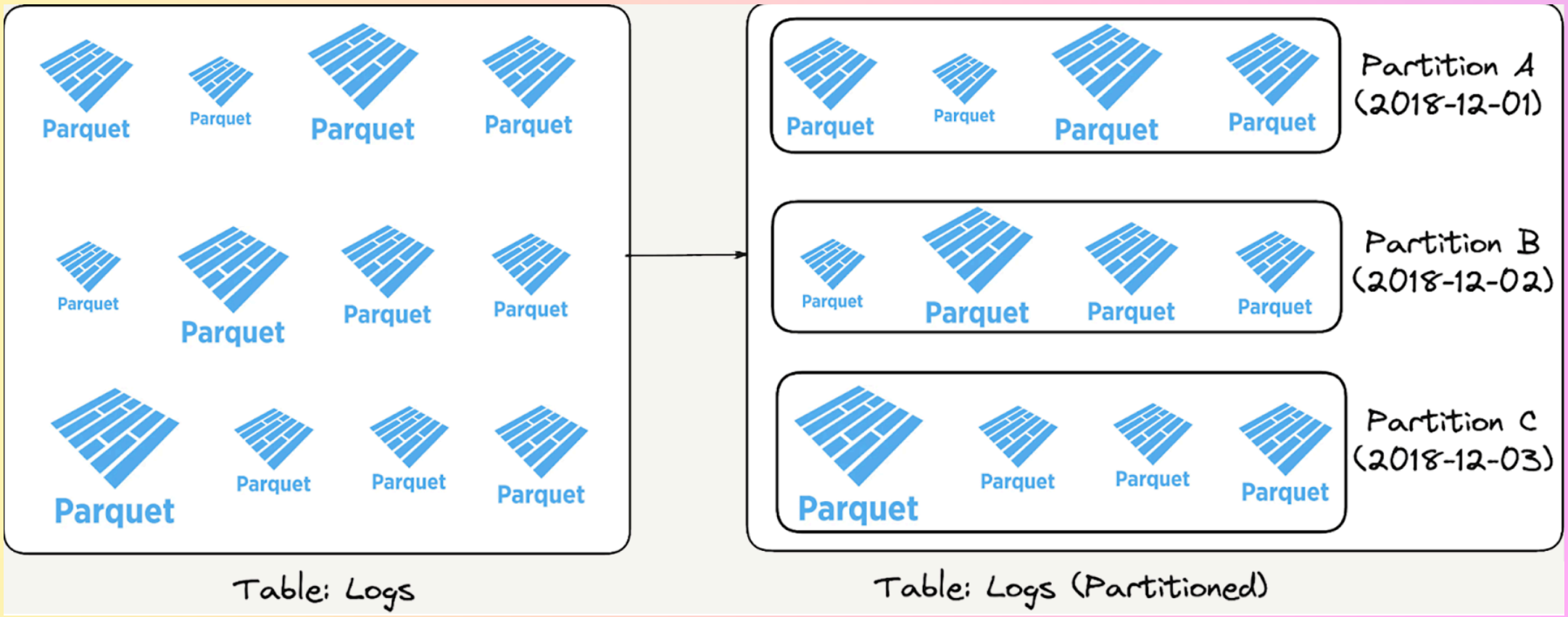
02 Compaction

03 Clustering

04 Data Skipping

05 Cleaning

Partitioning

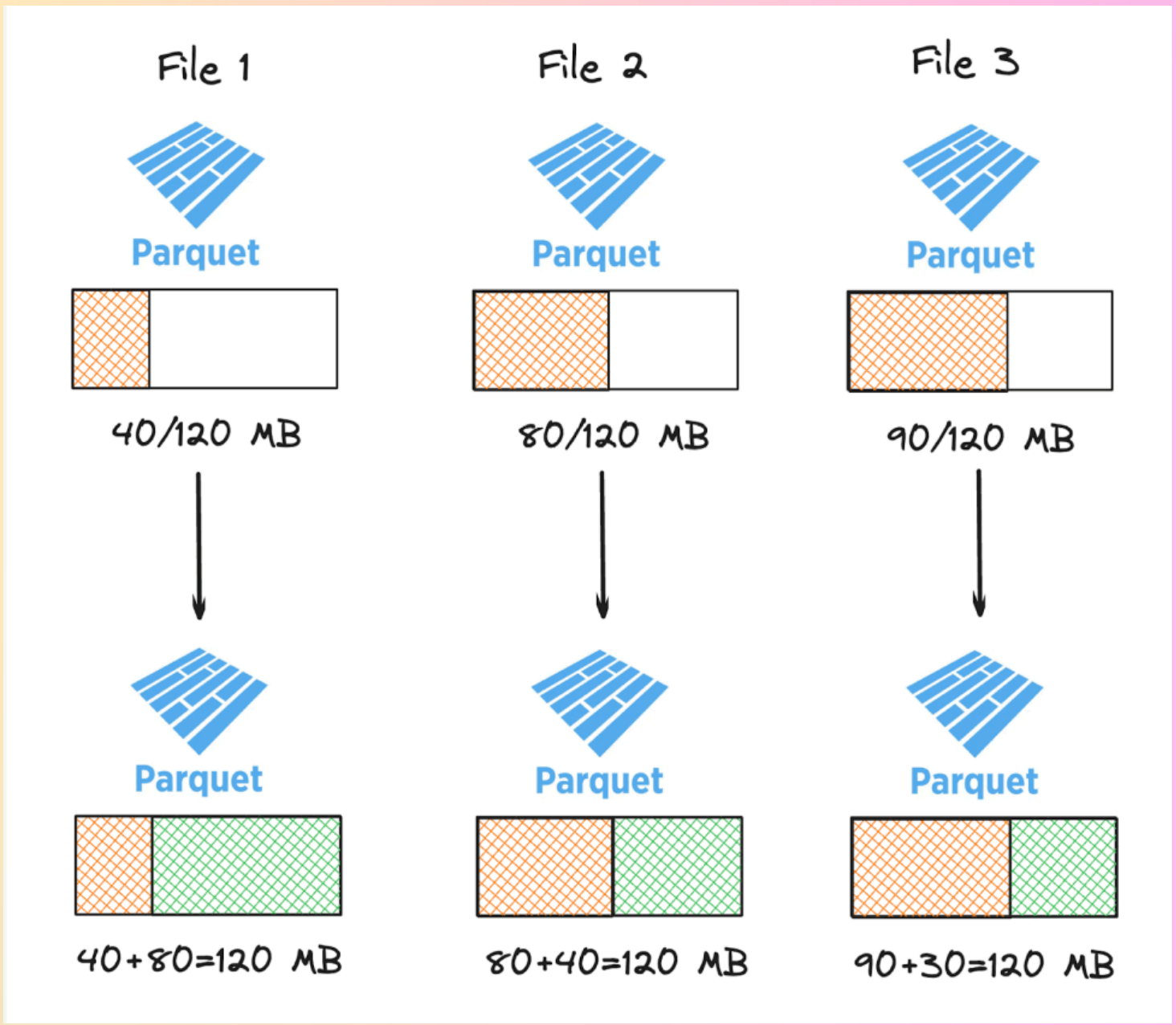


Group similar rows together when writing.

Retrieve data from only relevant partition.

Horizontal Scaling & Availability

Compaction



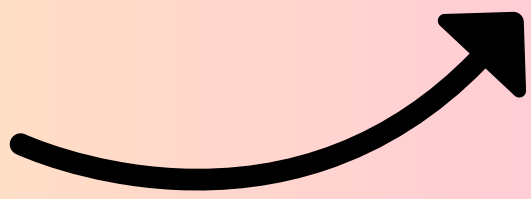
Deals with the small file problem.

Compact small files to an Optimal Size.

Techniques: Bin-packing.

Data Skipping

	Min	Max	Count
File01.parquet (date)	2022-06-01	2022-06-01	1500
File01.parquet (sales)	600	1400	1500
File02.parquet (date)	2022-07-01	2022-12-01	2000

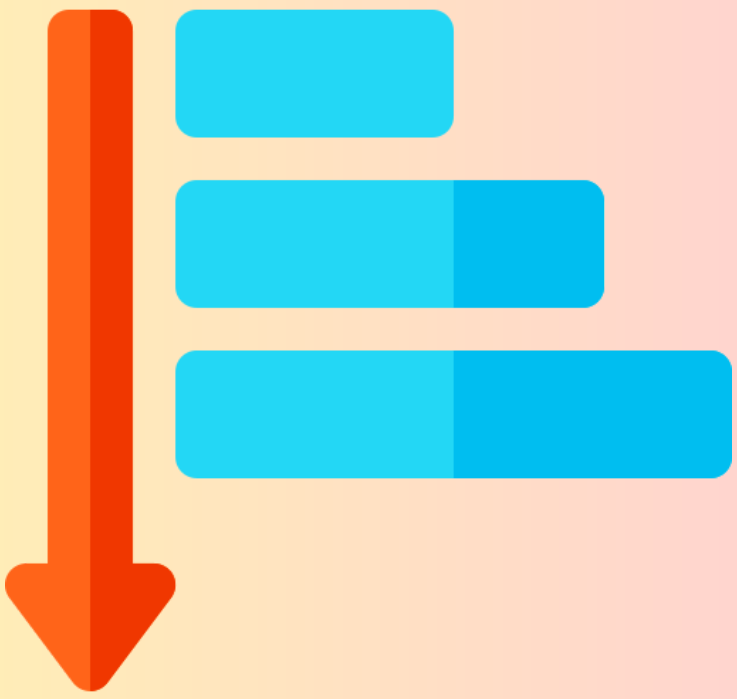


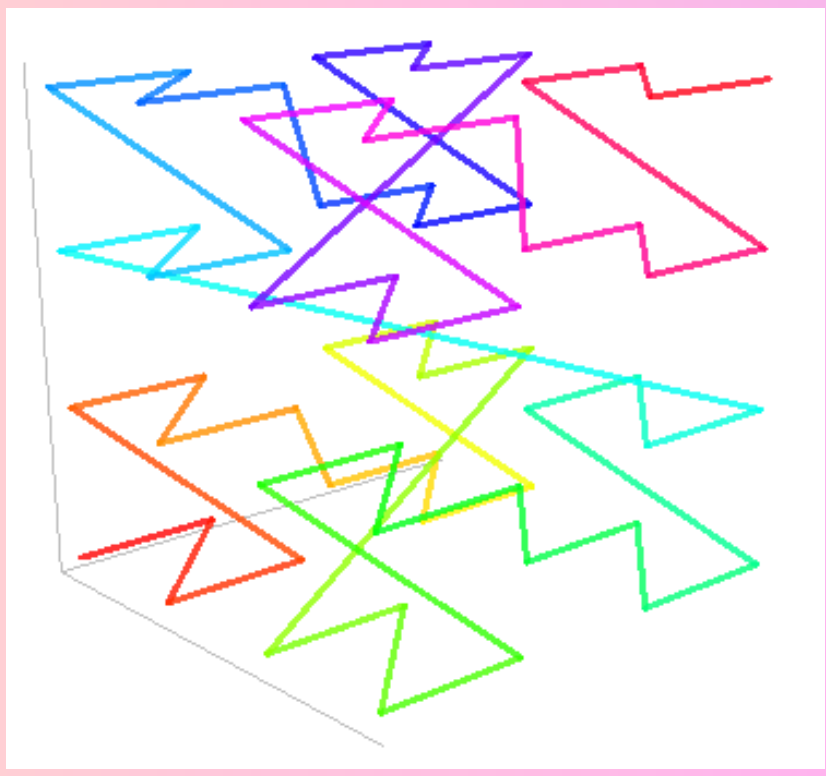
Uses Metrics from file formats like Parquet.

Min-Max Stats, Bloom Filters

Skip Irrelevant Data Files

Clustering





Reorganize & Group Data witin Files

Addresses misalignment between arrival & even time

Techniques: Linear Sort, Z-Order, Hilbert Curves

Cleaning

As data & metadata accumulates,
the storage becomes bloated.

Cleaning periodically removes
outdated snapshots.