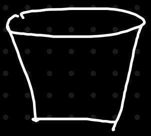
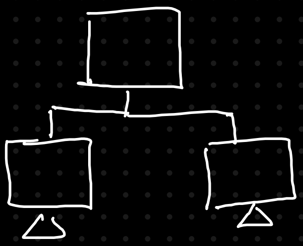
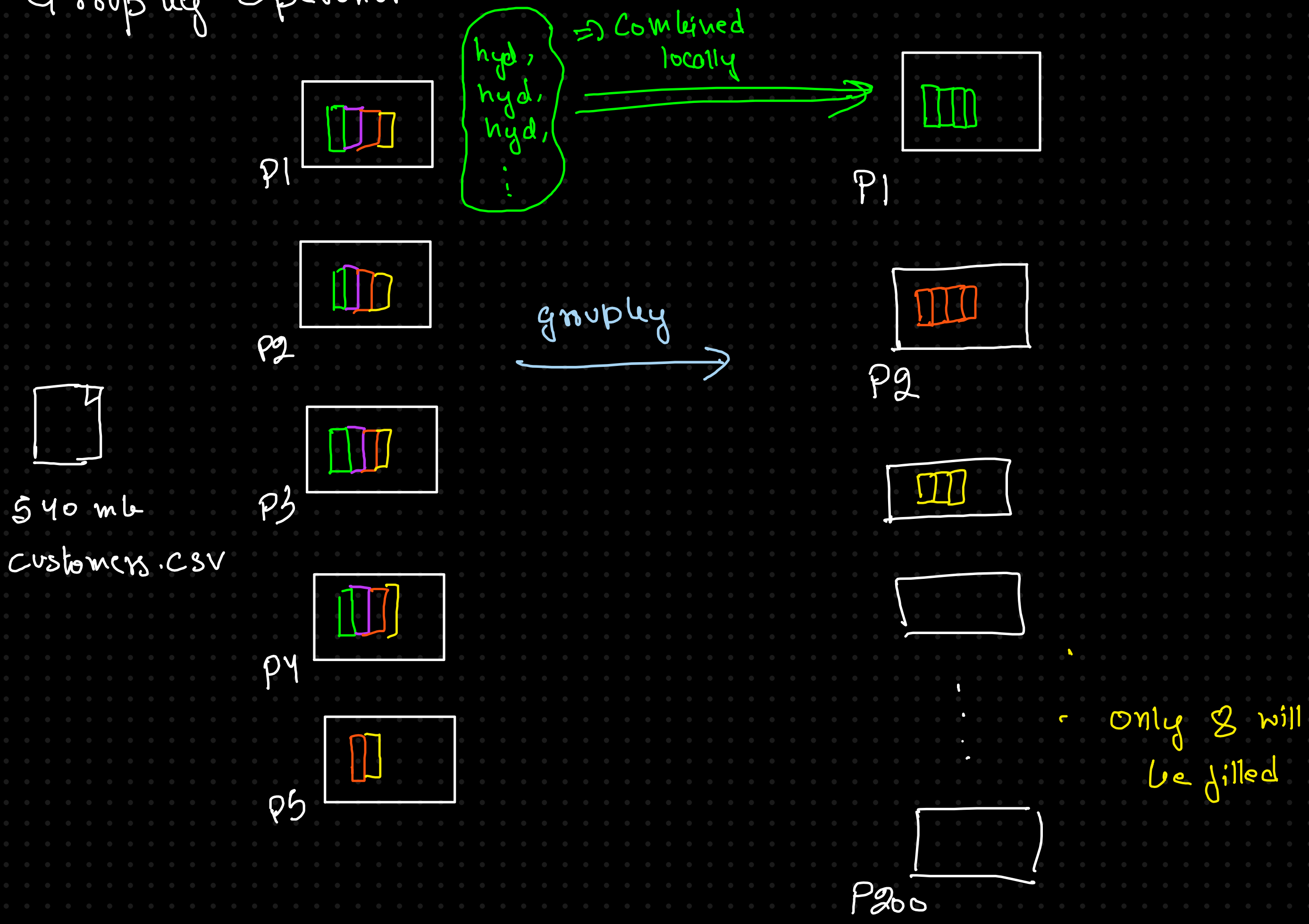
Buckets



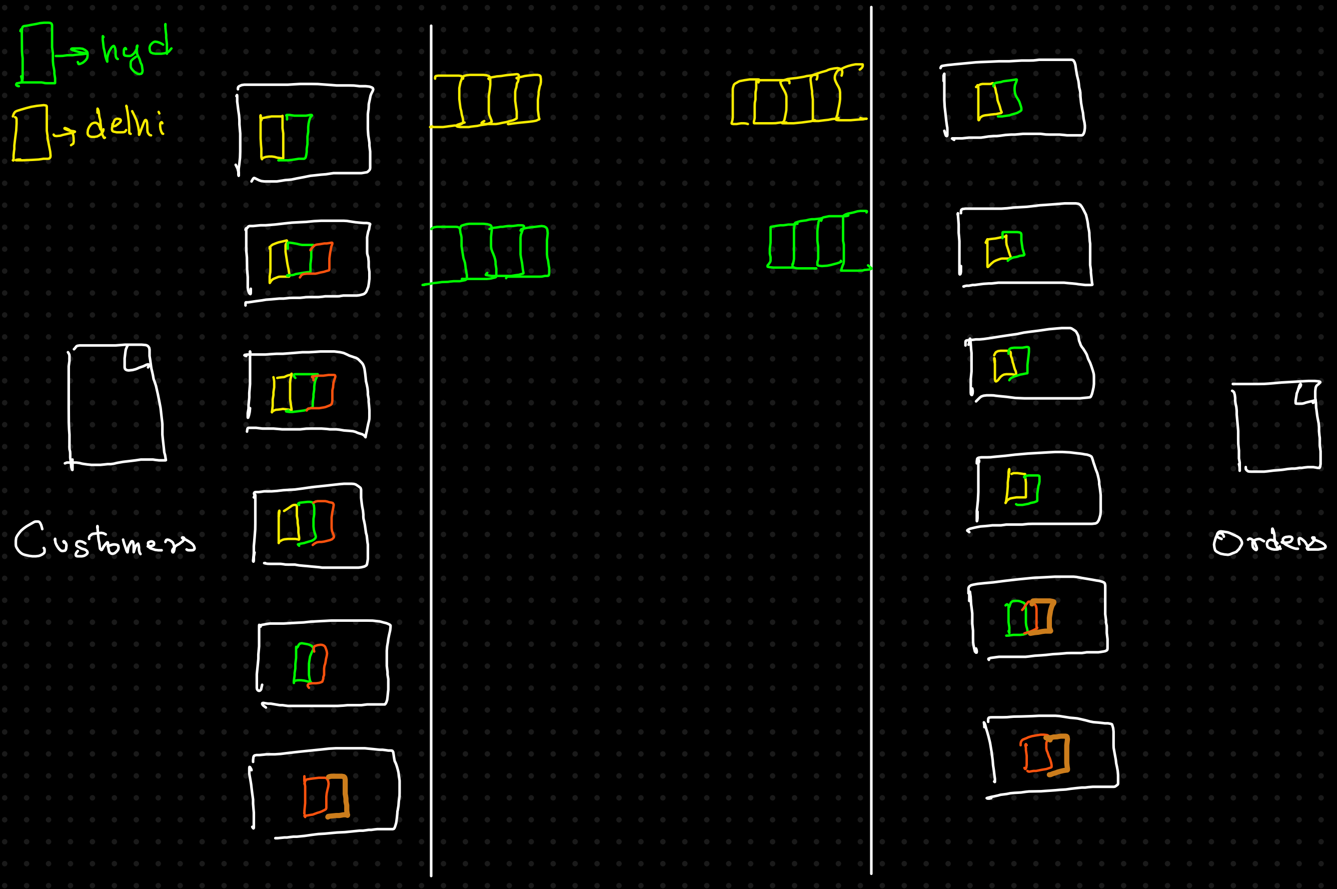


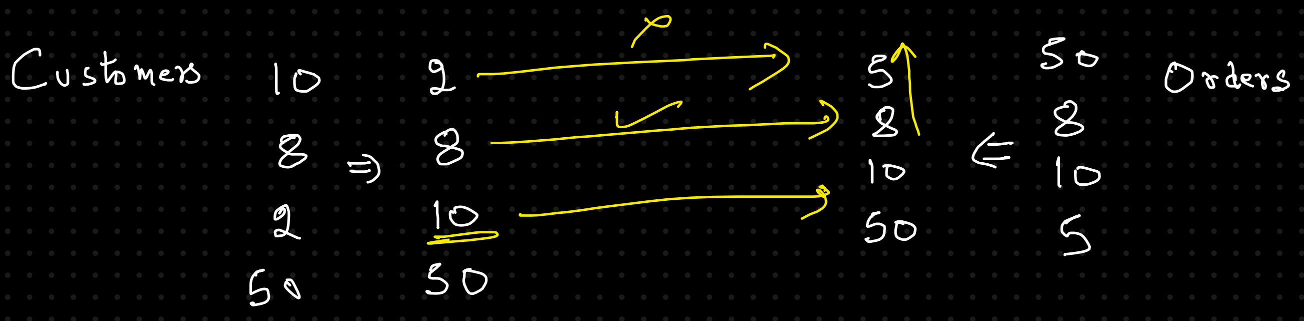
Groupby Operation



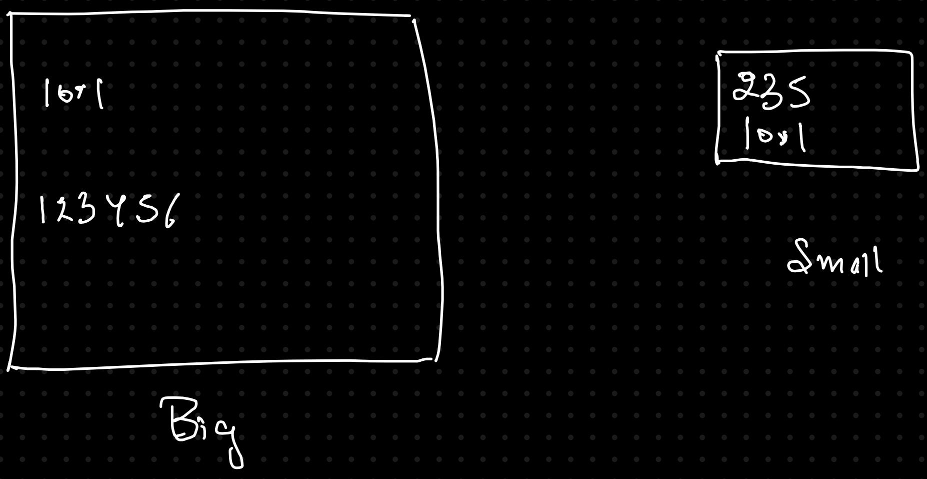
Join in Spark

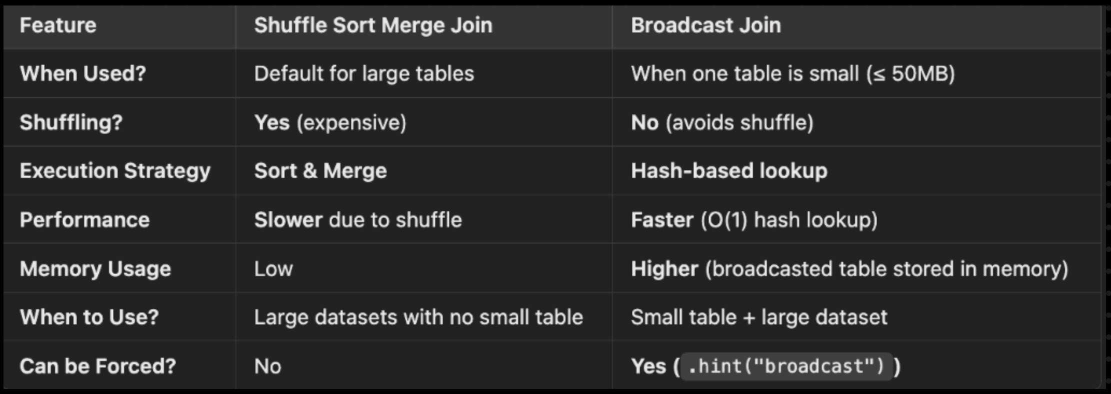
- Default
- Broadcast
- Skew





Broadcast Join





Feature	Shuffle Sort Merge Join	Broadcast Join
When Used?	Default for large tables	When one table is small (≤ 50MB)
Shuffling?	**Yes** (expensive)	**No** (avoids shuffle)
Execution Strategy	**Sort & Merge**	**Hash-based lookup**
Performance	**Slower** due to shuffle	**Faster** (O(1) hash lookup)
Memory Usage	Low	**Higher** (broadcasted table stored in memory)
When to Use?	Large datasets with no small table	Small table + large dataset
Can be Forced?	No	**Yes** (`.hint("broadcast")`)

Skewness in the Data