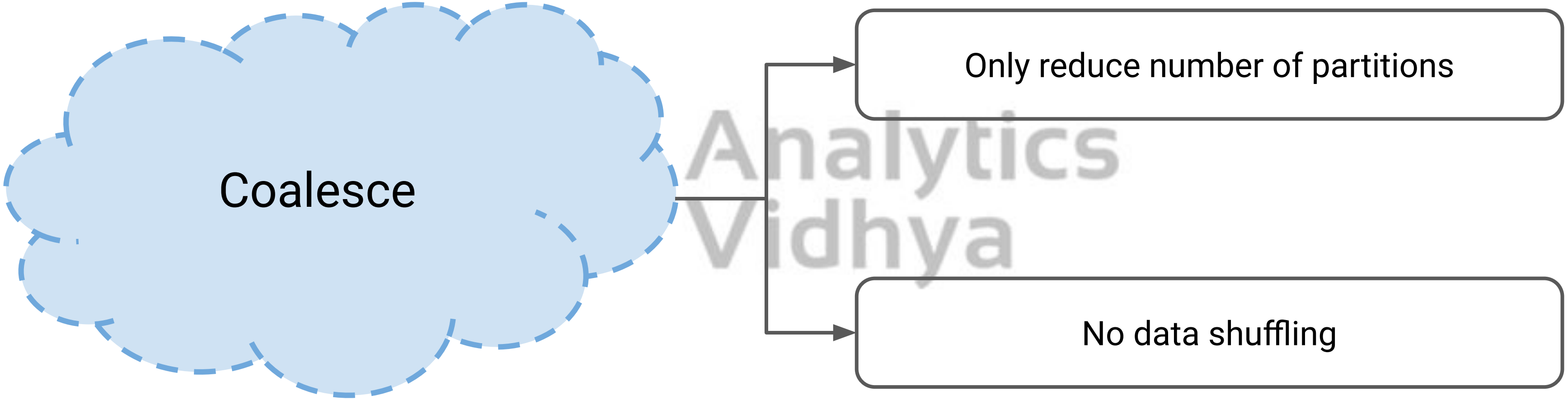
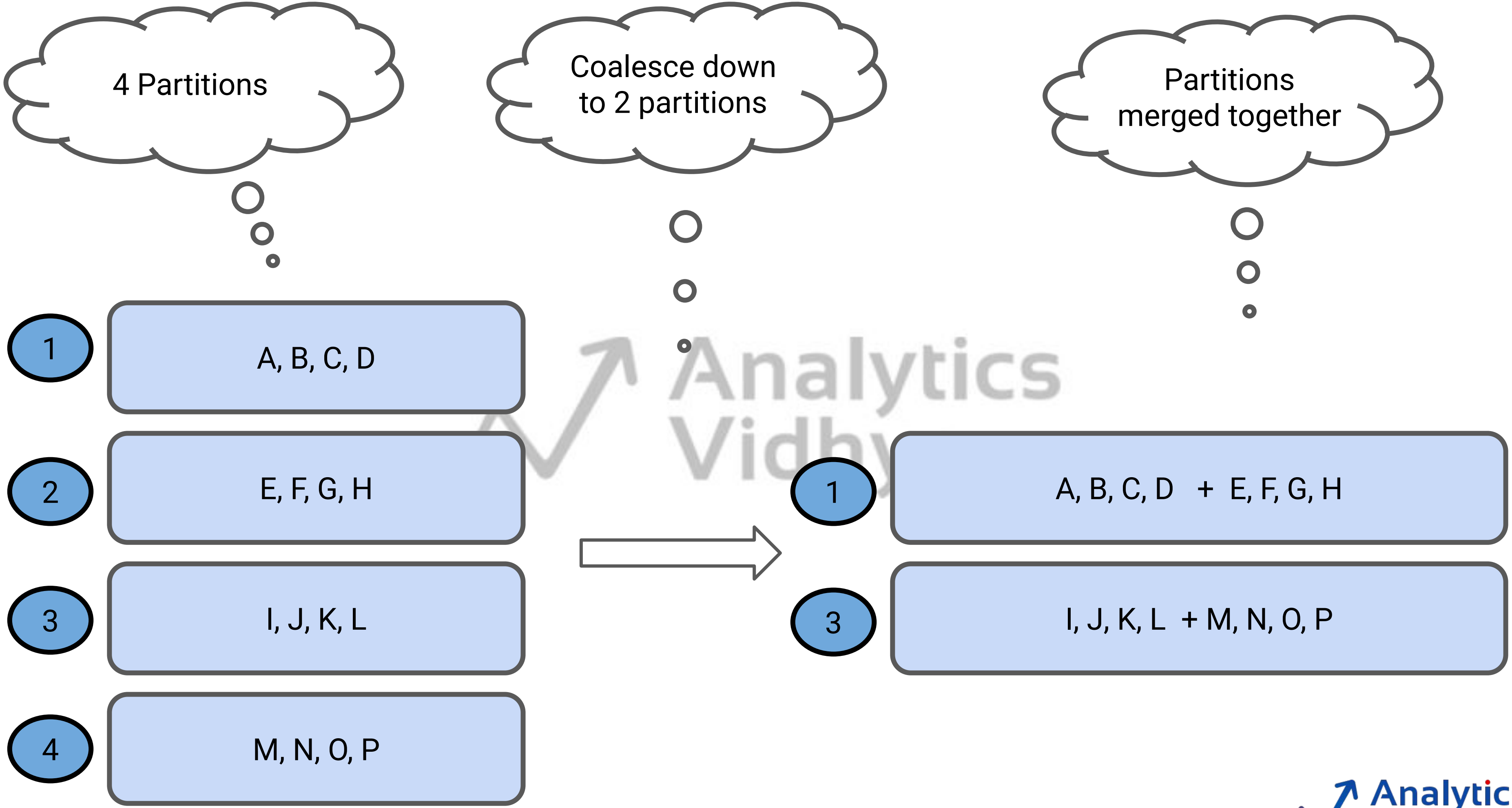
Coalesce vs Repartition



Coalesce

Only reduce number of partitions

No data shuffling



4 Partitions

Coalesce down to 2 partitions

Partitions merged together

1 A, B, C, D

2 E, F, G, H

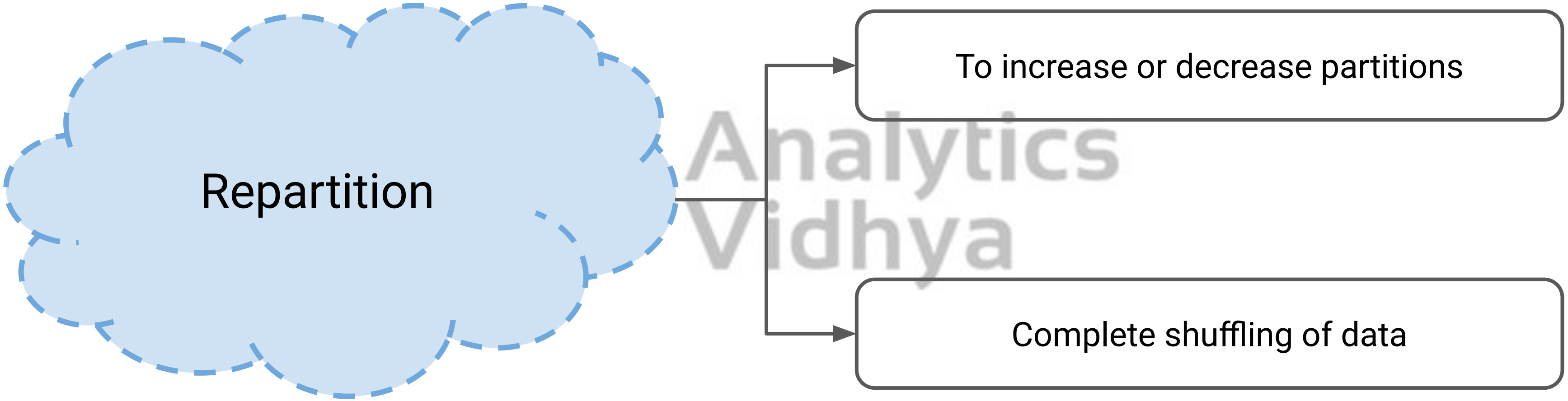
3 I, J, K, L

4 M, N, O, P

1 A, B, C, D + E, F, G, H

3 I, J, K, L + M, N, O, P

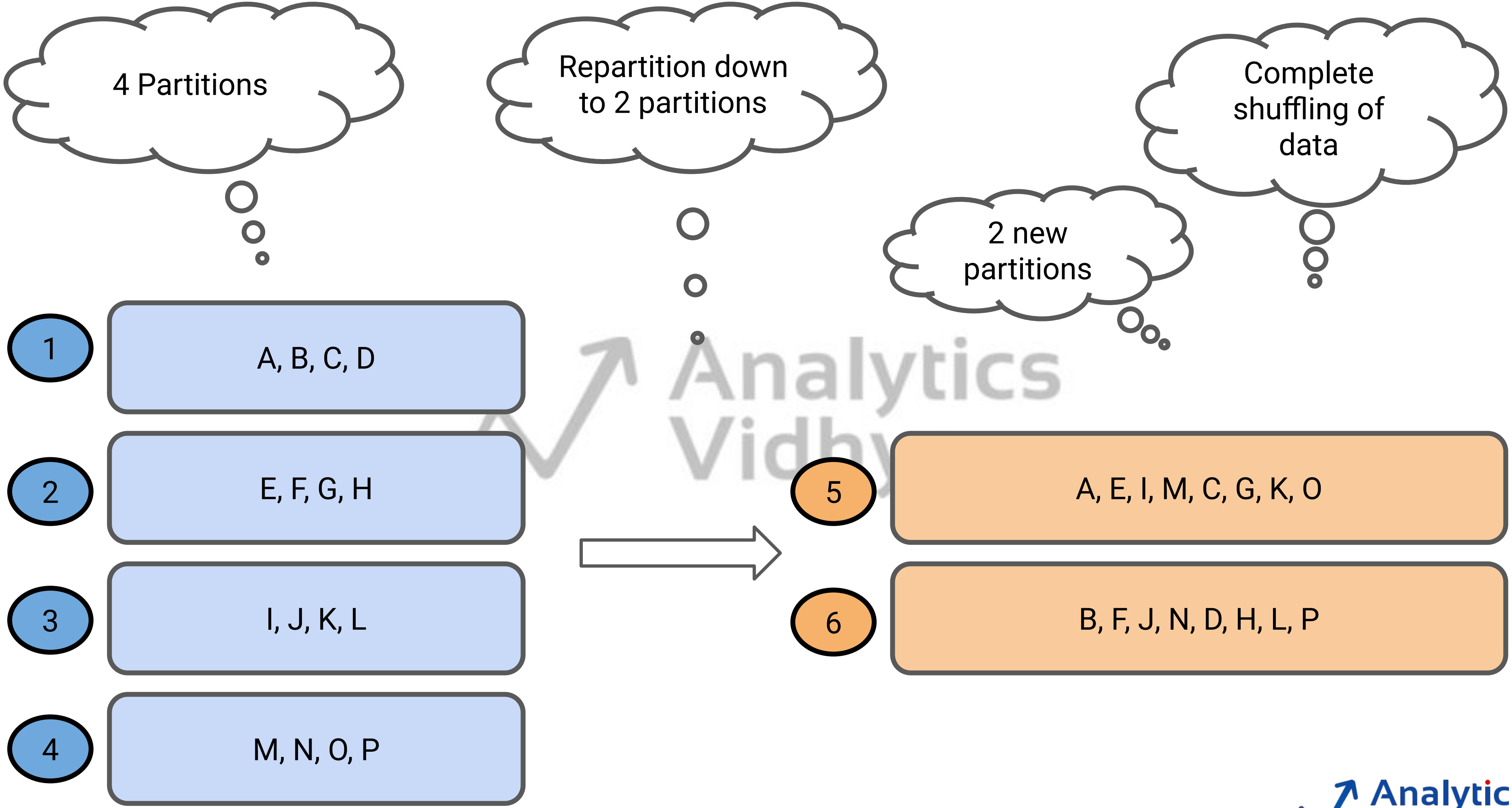
Repartition



Repartition

To increase or decrease partitions

Complete shuffling of data



4 Partitions

Repartition down to 2 partitions

Complete shuffling of data

2 new partitions

1 A, B, C, D

2 E, F, G, H

3 I, J, K, L

4 M, N, O, P

5 A, E, I, M, C, G, K, O

6 B, F, J, N, D, H, L, P

Partition tuning: Repartition or Coalesce ?

If your dataset is **skewed**: use **repartition**

If you want **more partitions**: use **repartition**

If you want to drastically **reduce** the number of partitions(e.g. numPartitions =1): use **repartition**

If your dataset is **well balanced**(i.e. not skewed) and you want **fewer partitions,** but not dramatically fewer): use **coalesce**

Thank You