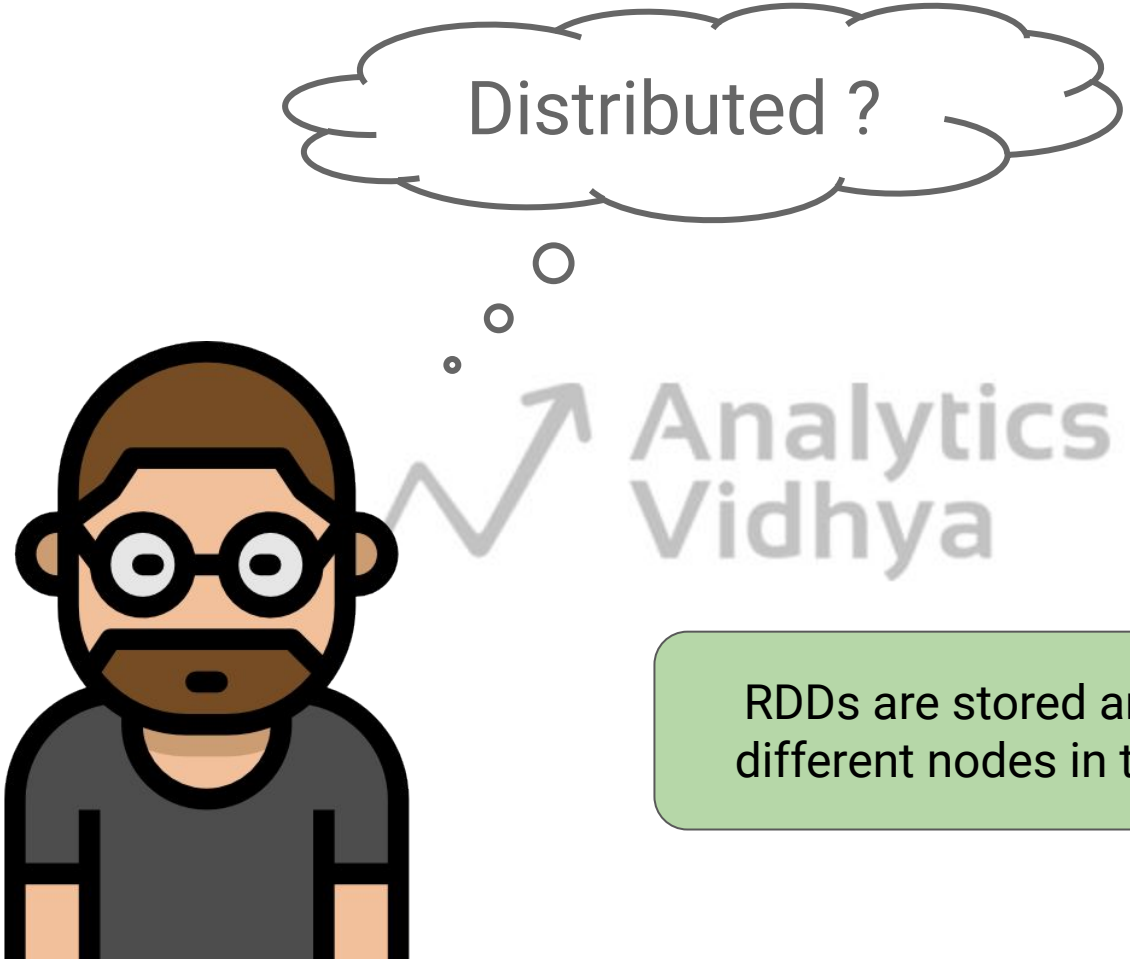
What are RDDs?

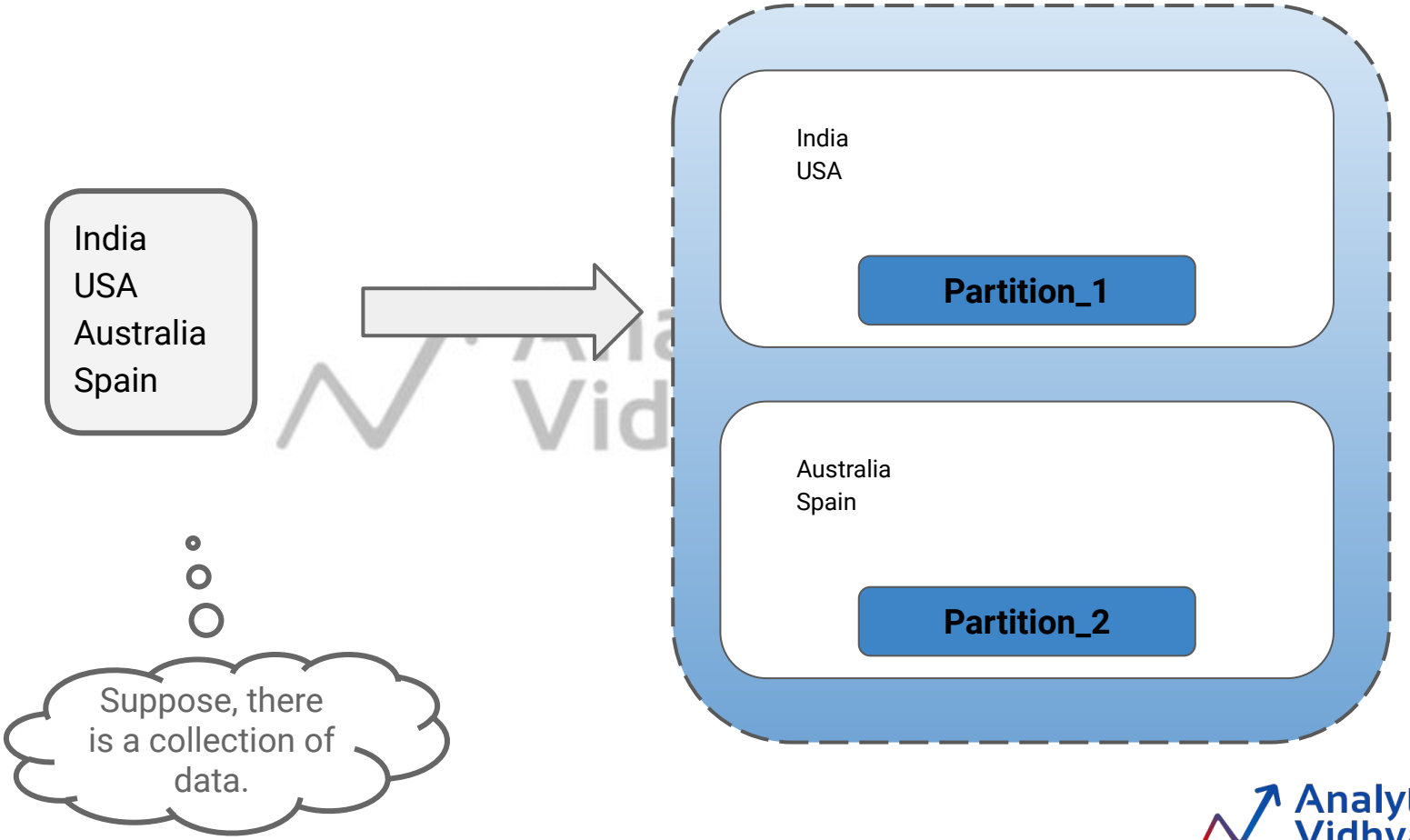
What are RDDs?

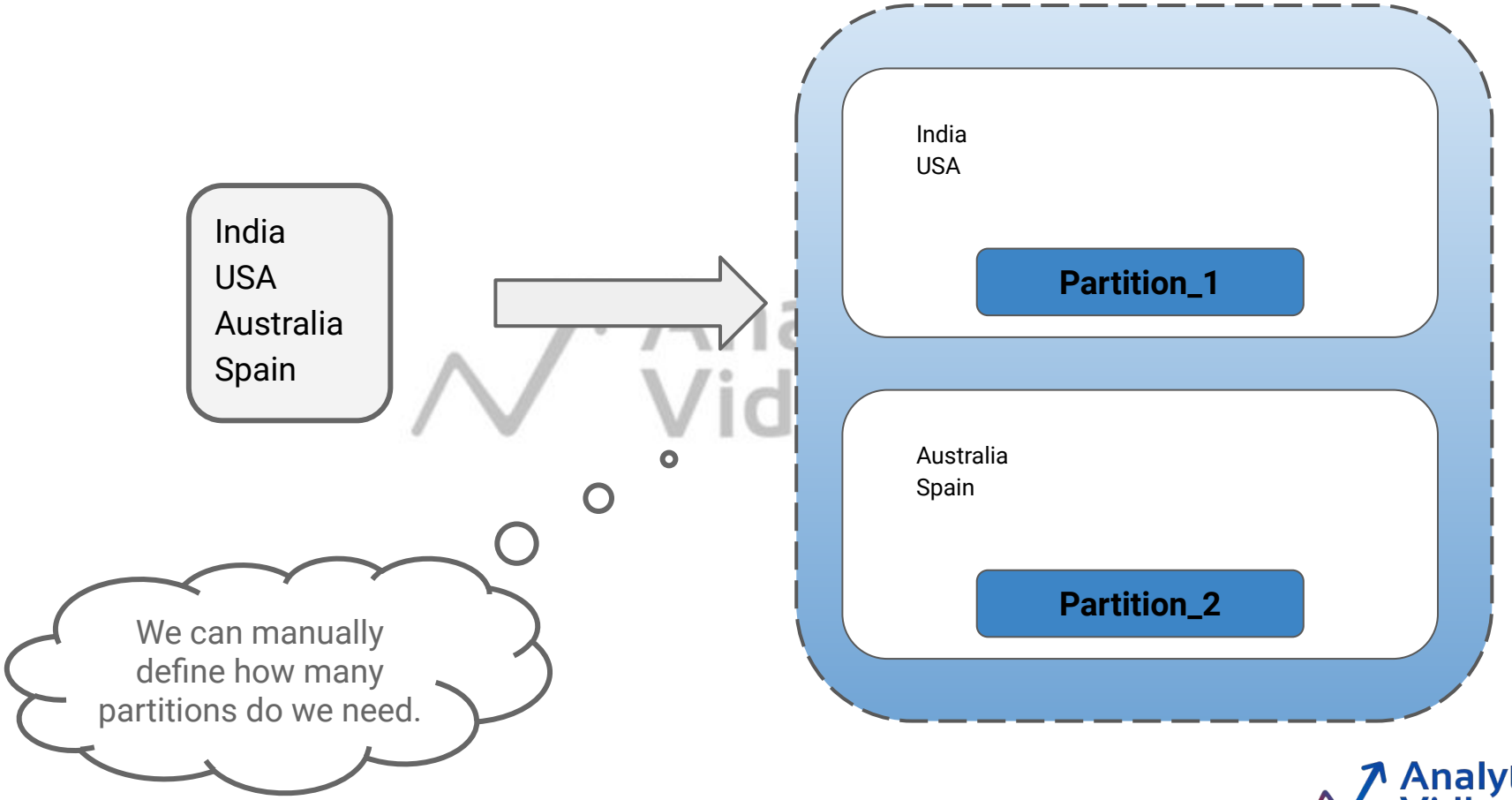
RDDs (Resilient Distributed Dataset) is the Spark's main abstraction and has the following features:

- Distributed
- Immutable
- Resilient

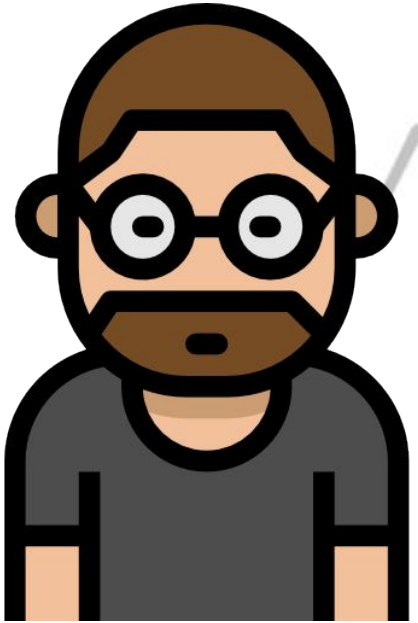


RDDs are stored and processed on different nodes in the Spark cluster.

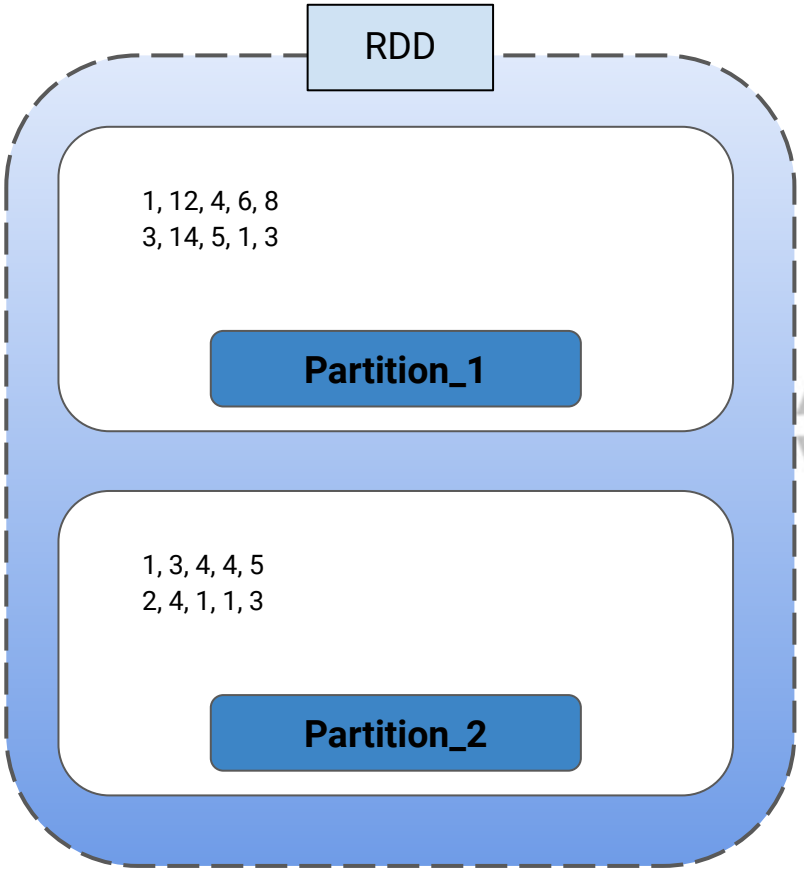




Immutable ?



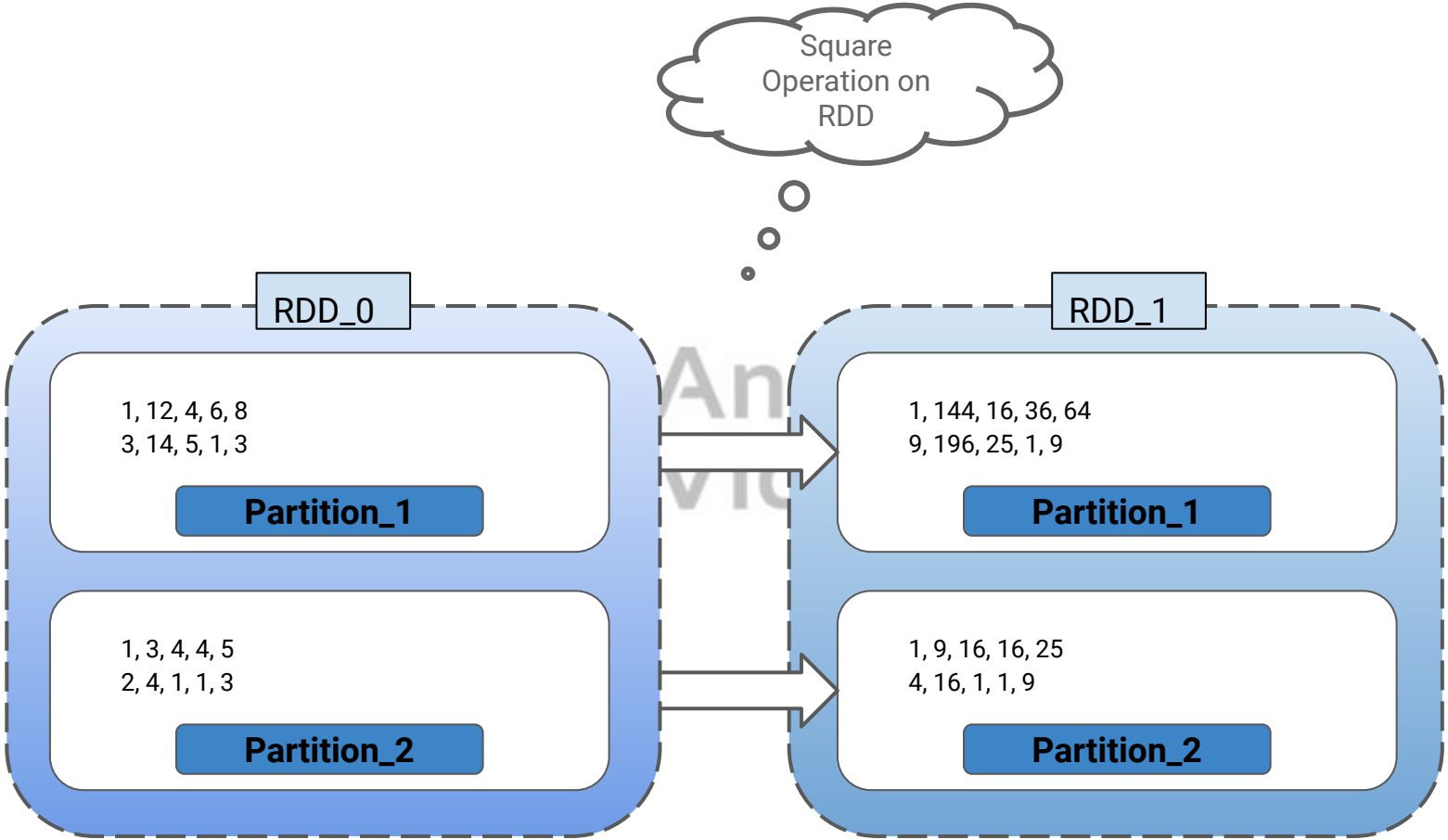
Once an RDD is created, it cannot be modified further. It can only be used to create a new RDD

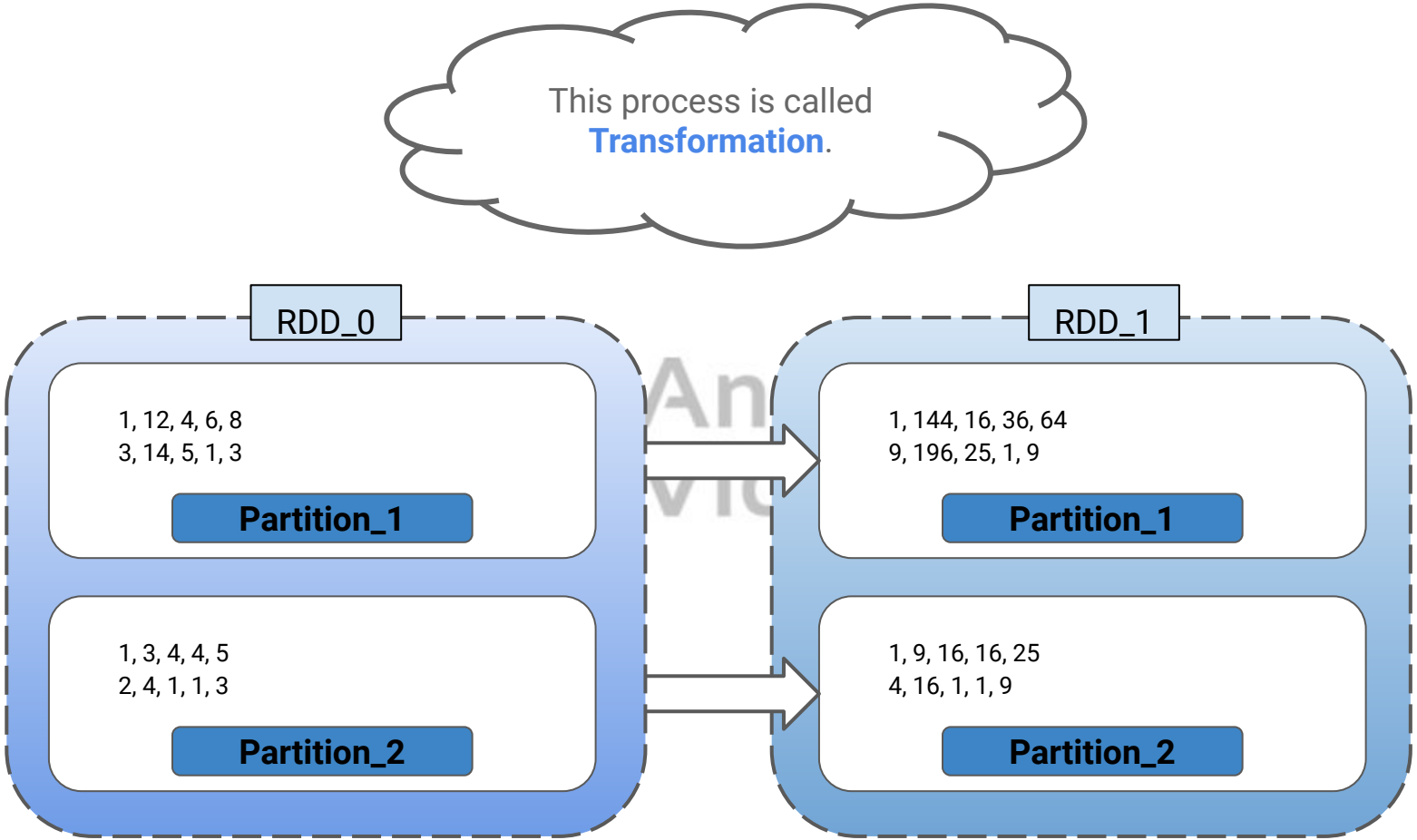


Now, if you want to do any operations on a RDD

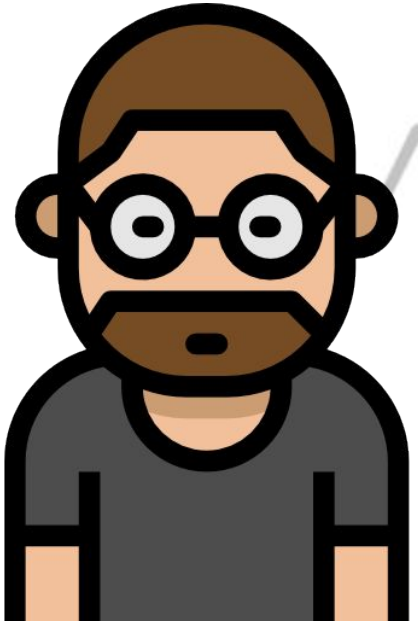
square of each number present in the data

It will **not update** within the same RDD rather it will **create another one**





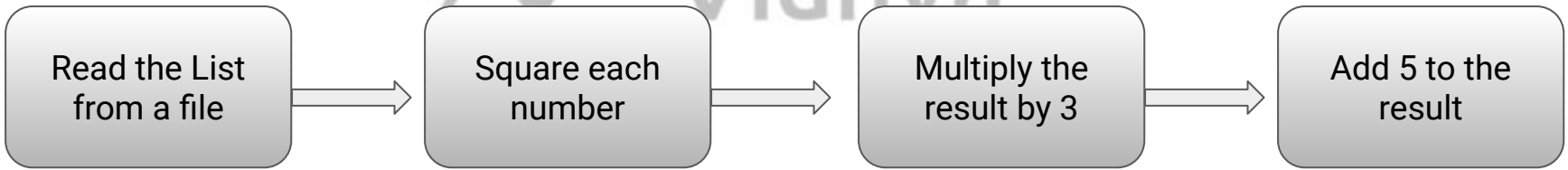
Resilient ?

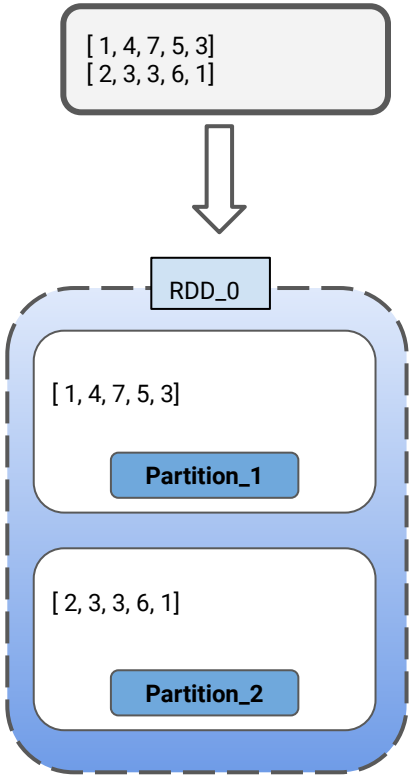


RDDs have the ability to continue the operations even if there is a failure.

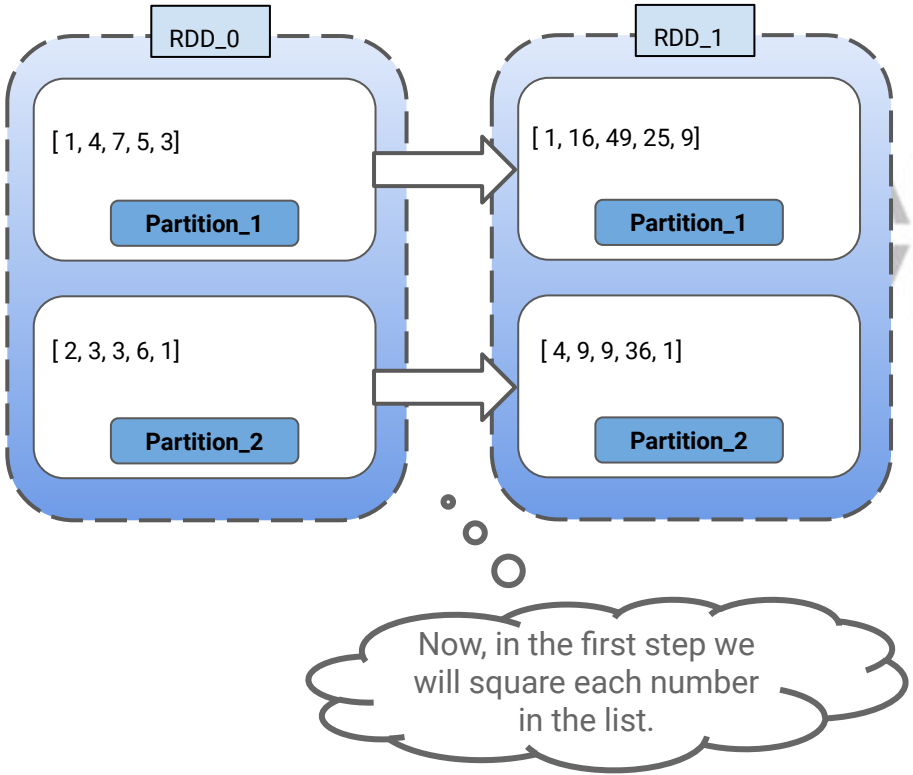
Let us Assume?

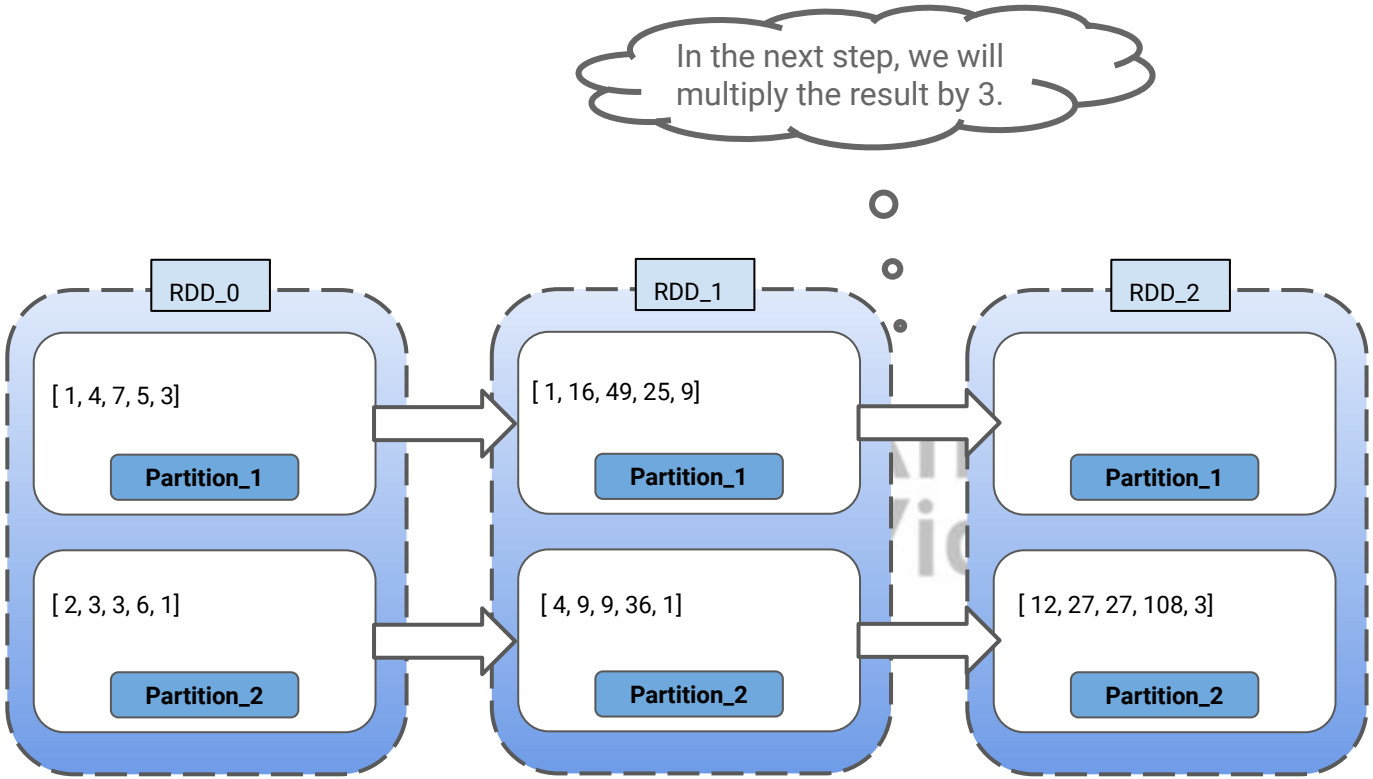
- We have to build the following operations on a list of numbers.

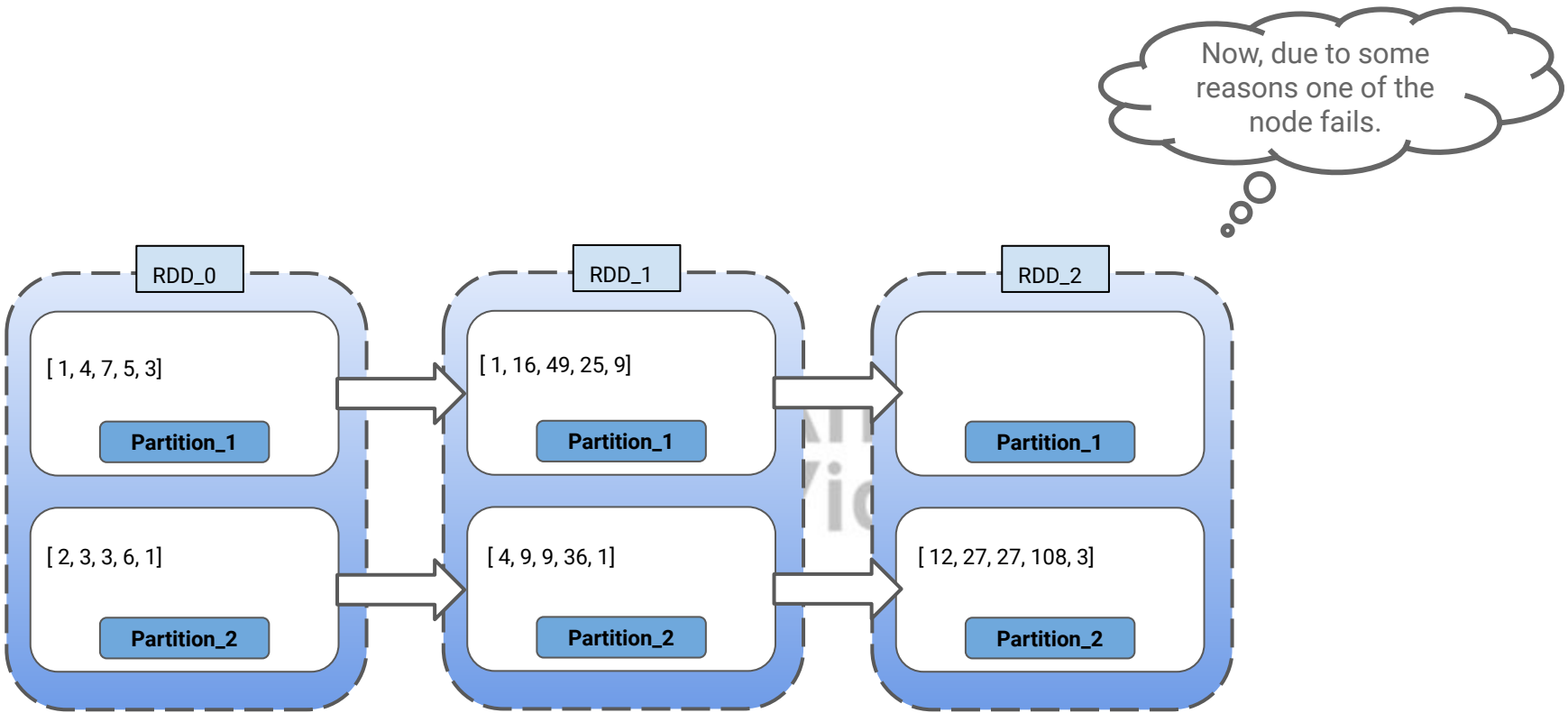


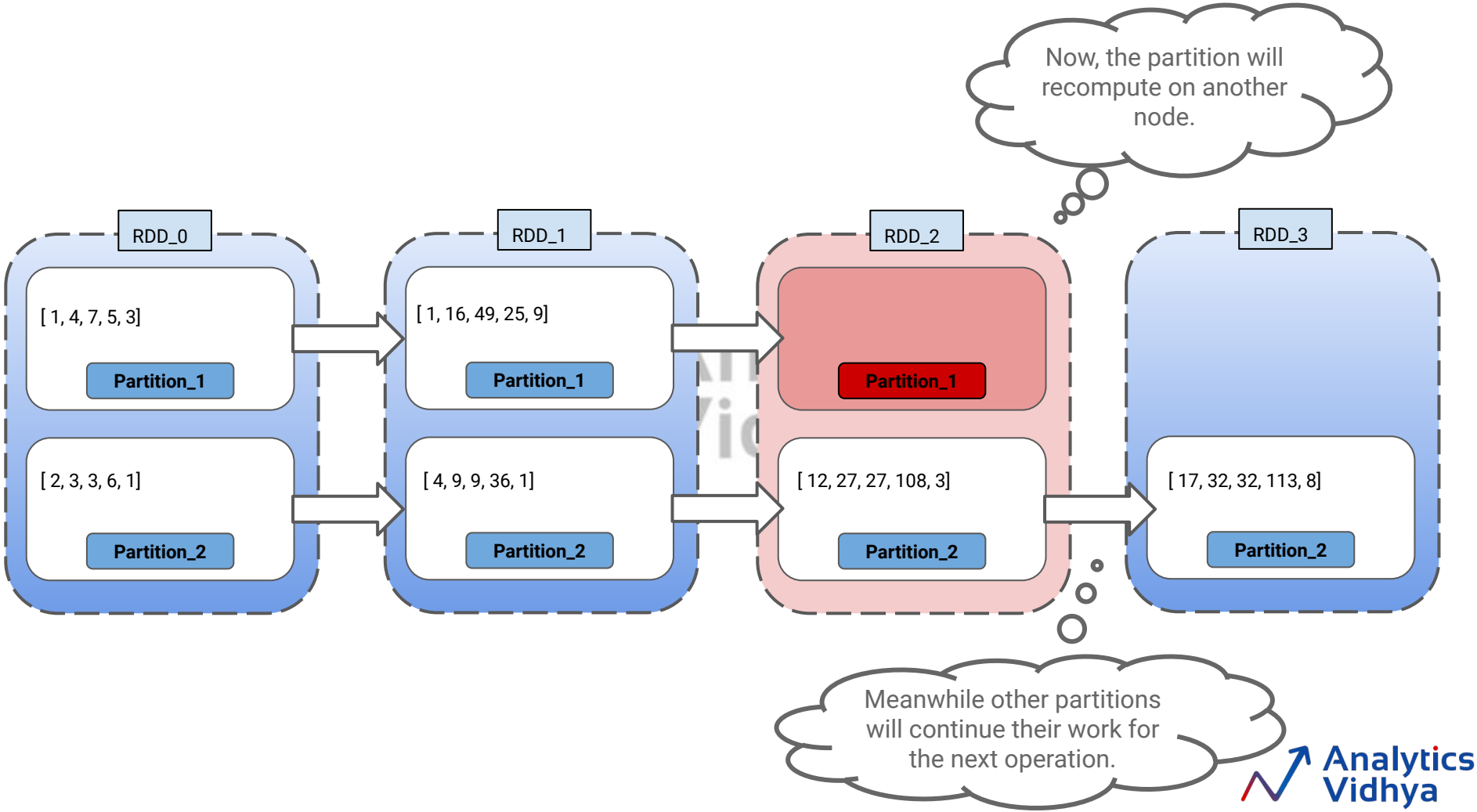


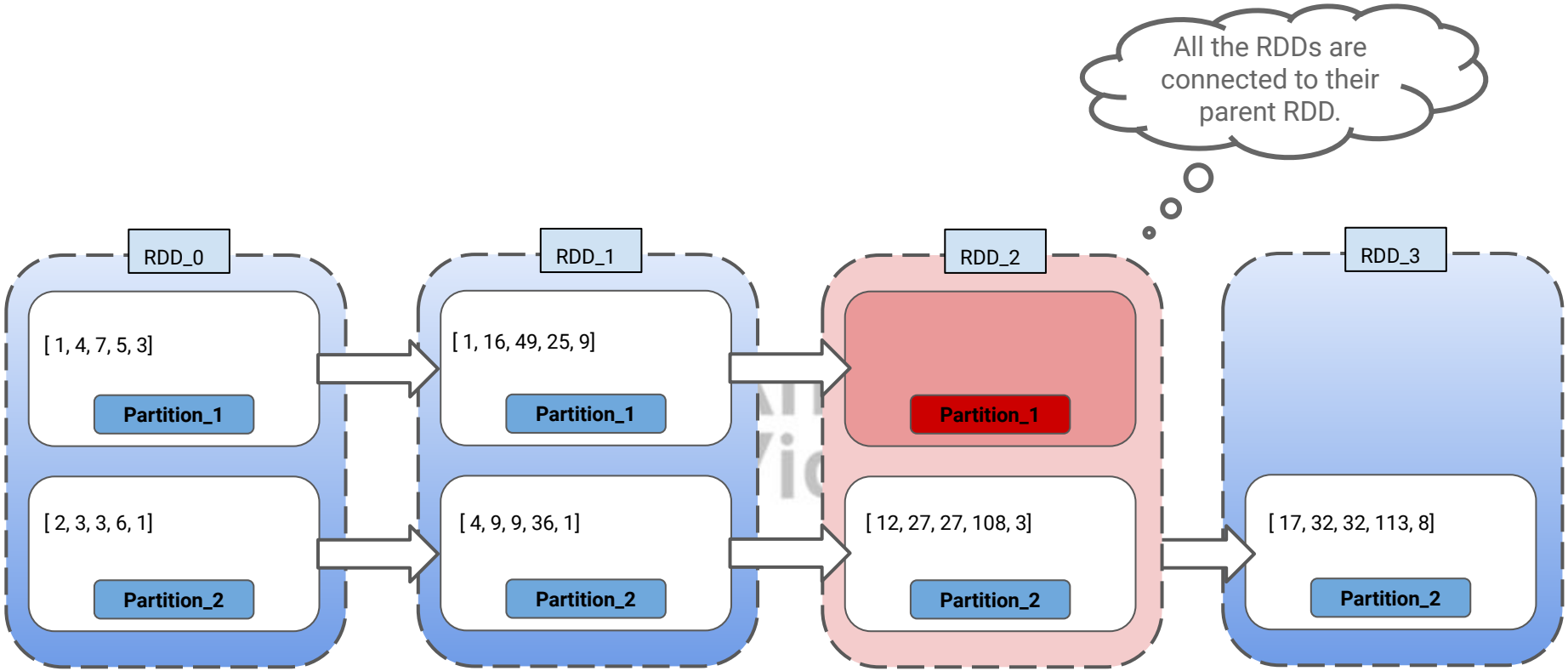
data is partitioned on 2 nodes.

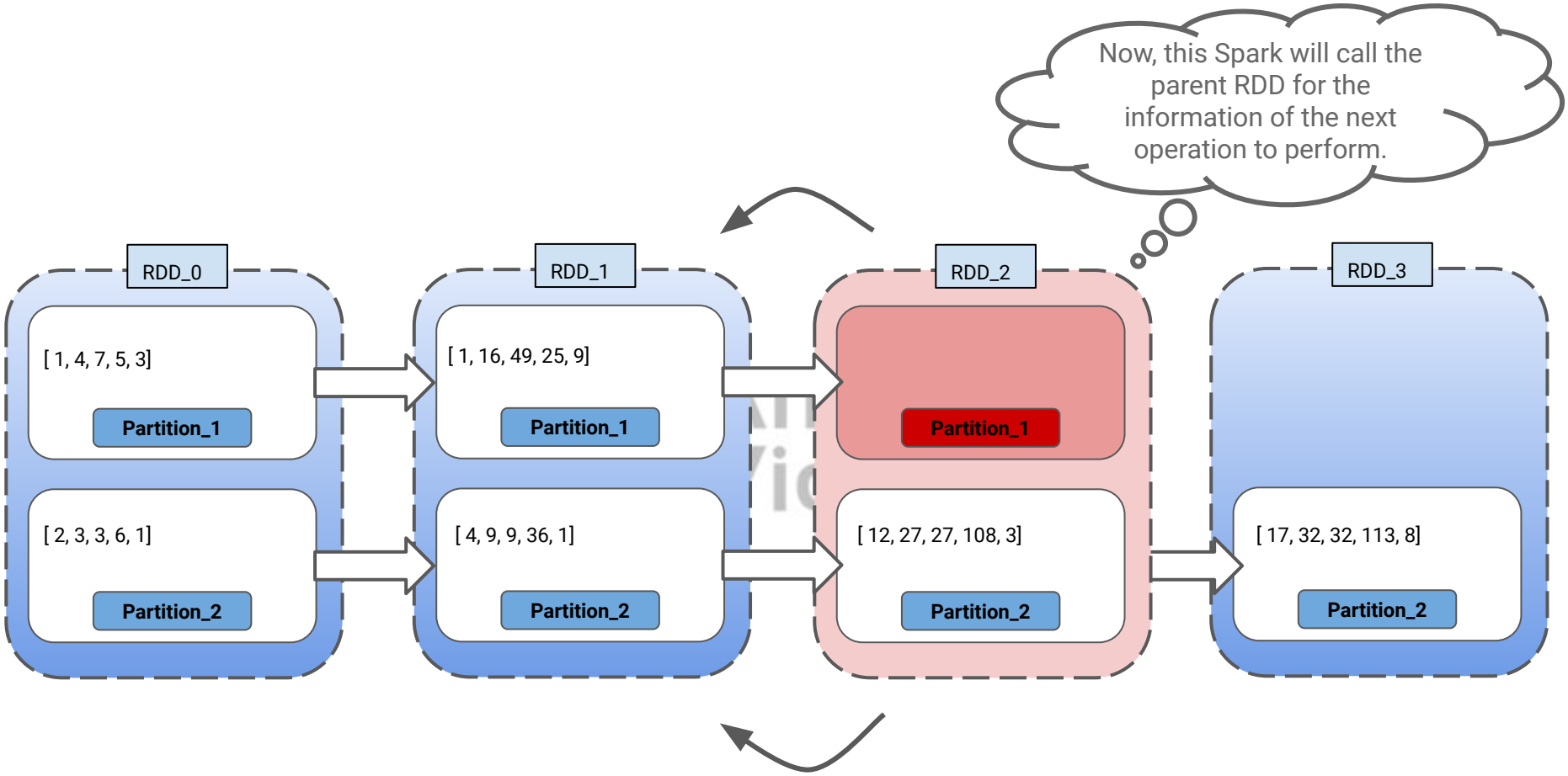


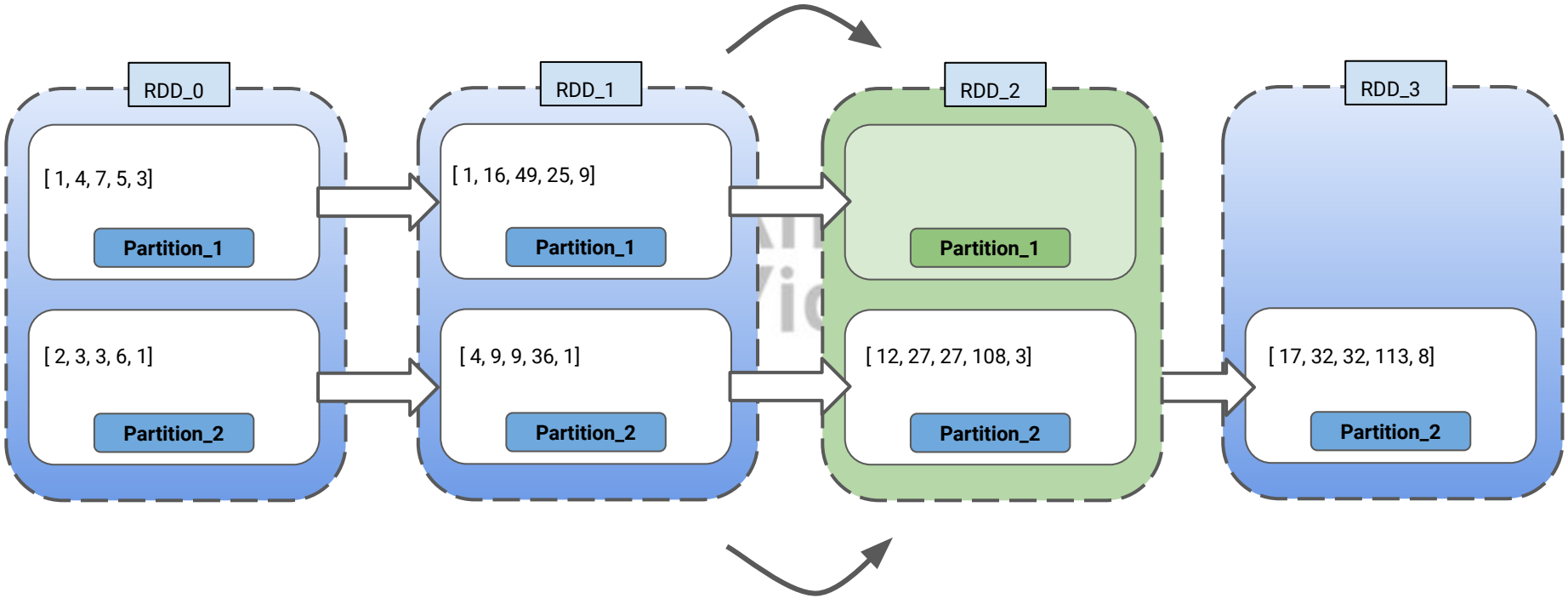


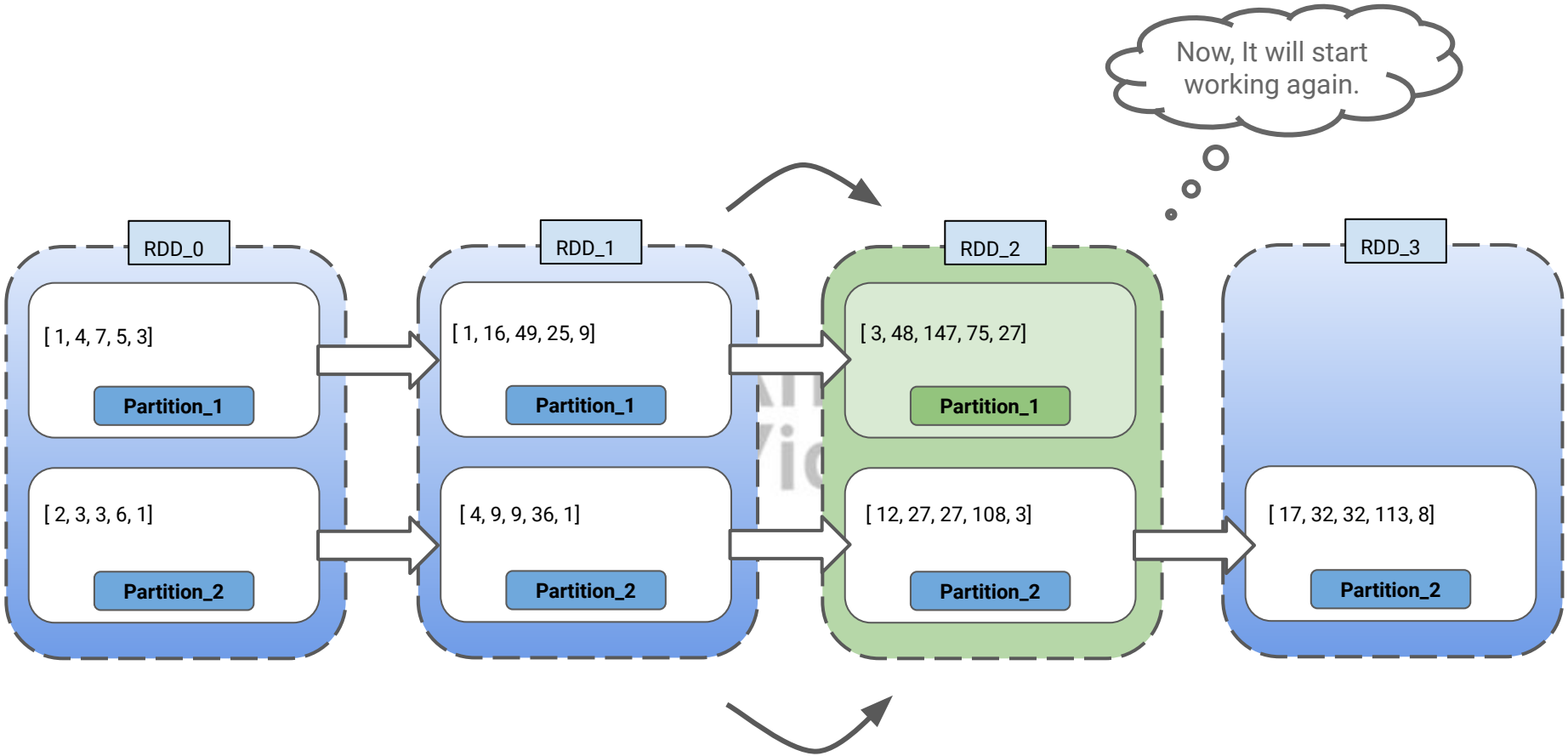


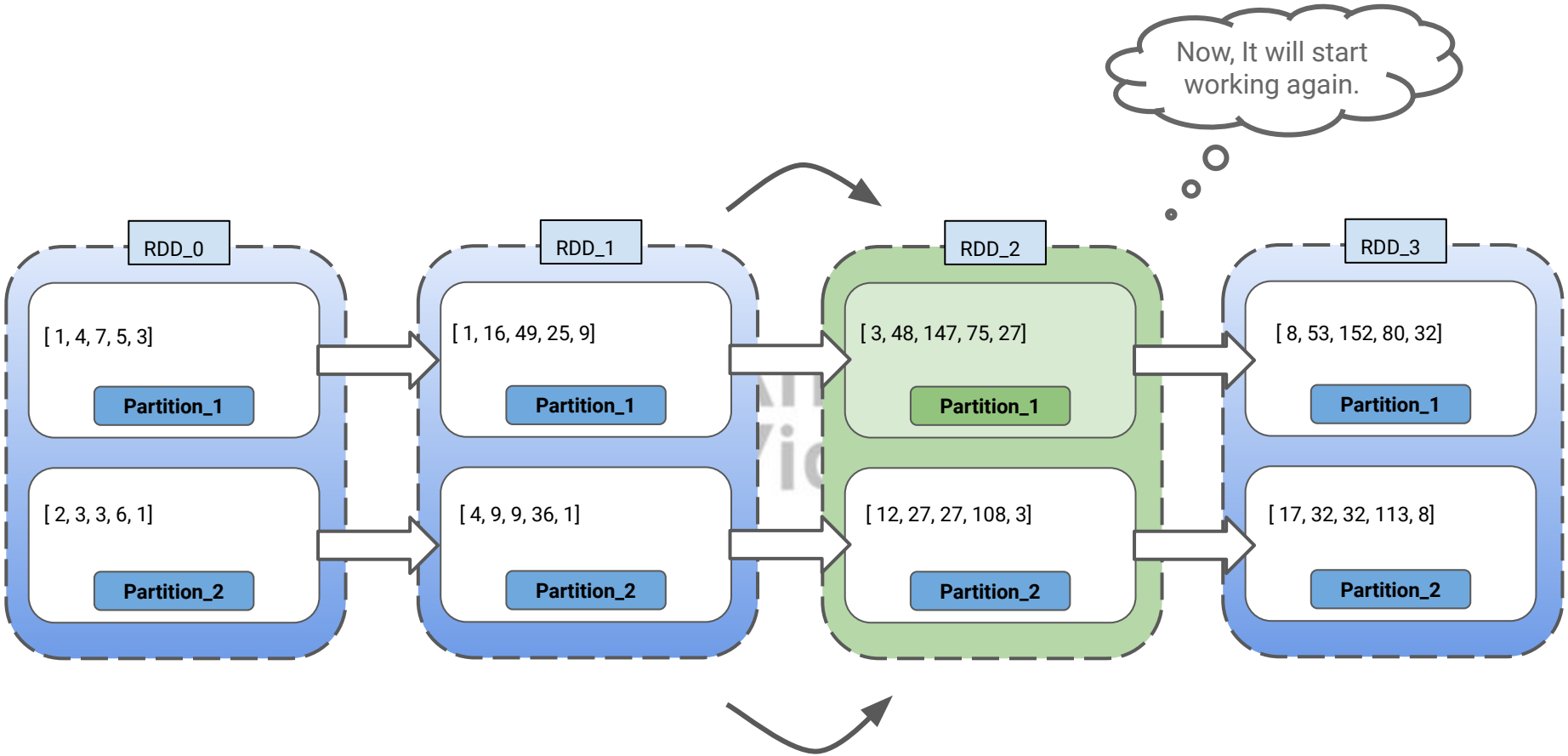

Now, the partition will recompute on another node.
RDD_0
[1, 4, 7, 5, 3]
Partition_1
[2, 3, 3, 6, 1]
Partition_2
RDD_1
[1, 16, 49, 25, 9]
Partition_1
[4, 9, 9, 36, 1]
Partition_2
RDD_2
Partition_1
[12, 27, 27, 108, 3]
Partition_2
RDD_3
[17, 32, 32, 113, 8]
Partition_2
Meanwhile other partitions will continue their work for the next operation.
Analytics Vidhya

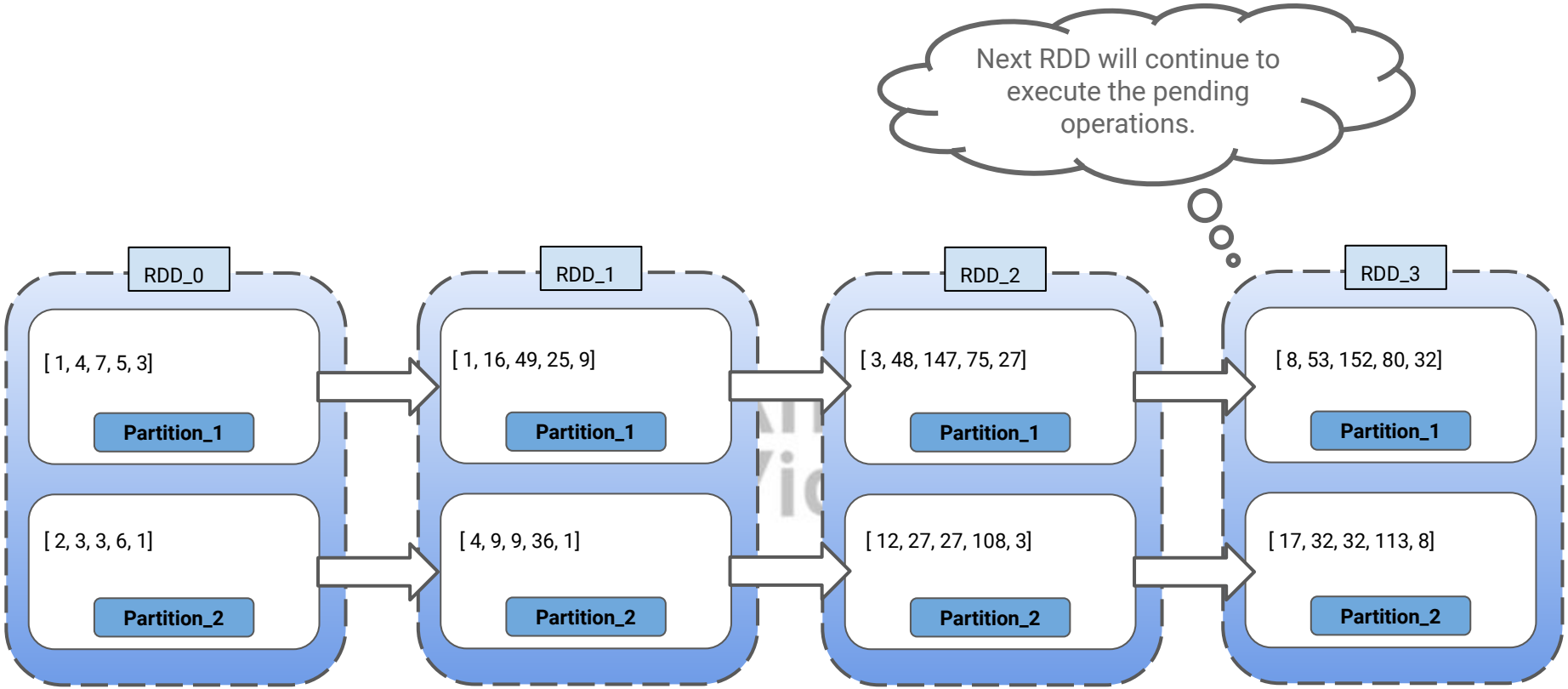












Thank You!!

Analytics
Vidhya