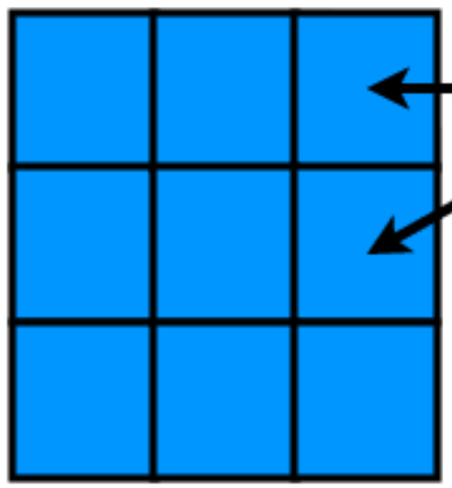
Bob

HDFS

DN1 DN2 DN3 DN4 DN5 DN6 ... DNn

HDFS

horizontal partitions



HDFS blocks
64MB (default)

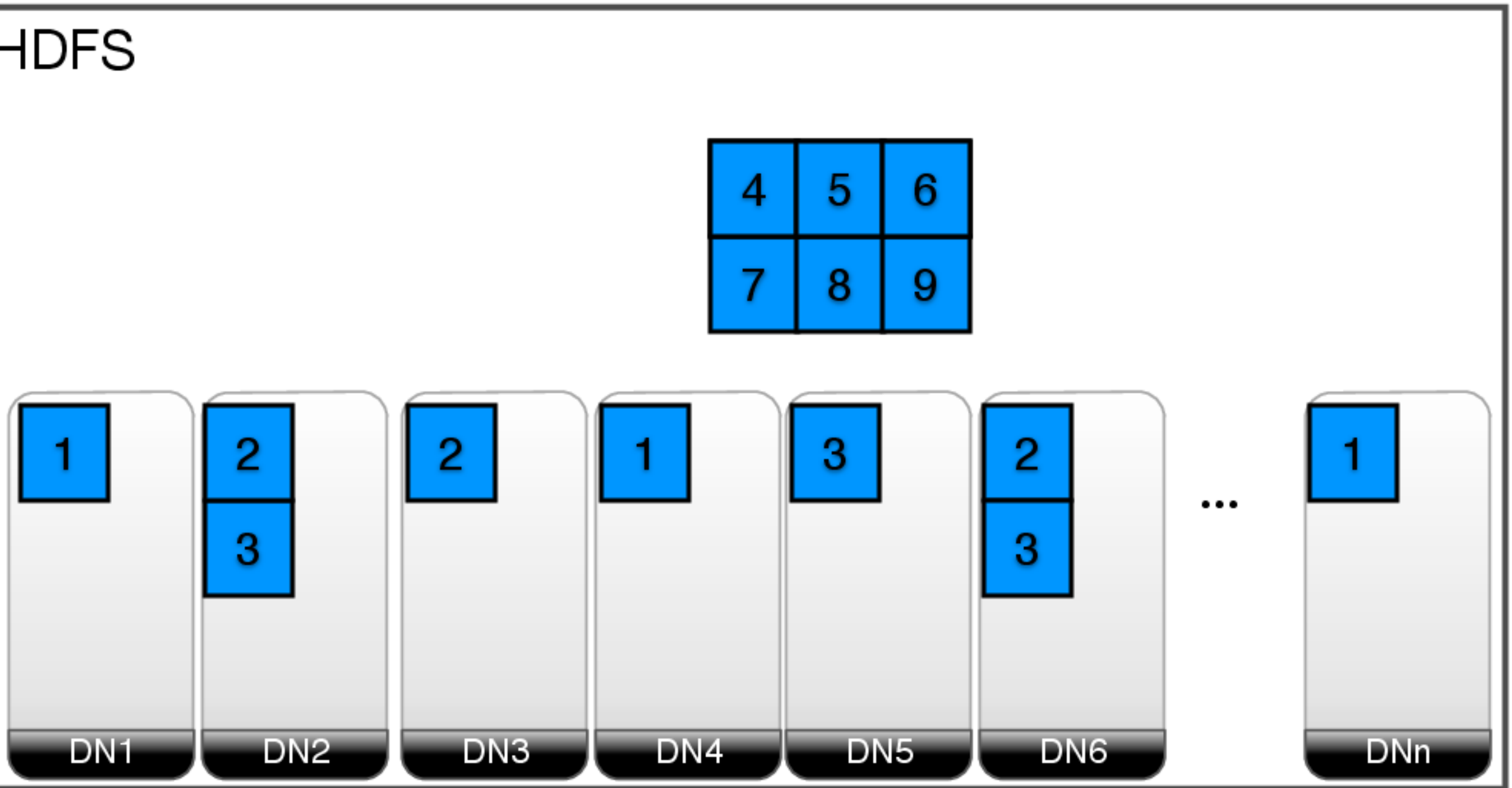
DN1 DN2 DN3 DN4 DN5 DN6 ... DNn

HDFS

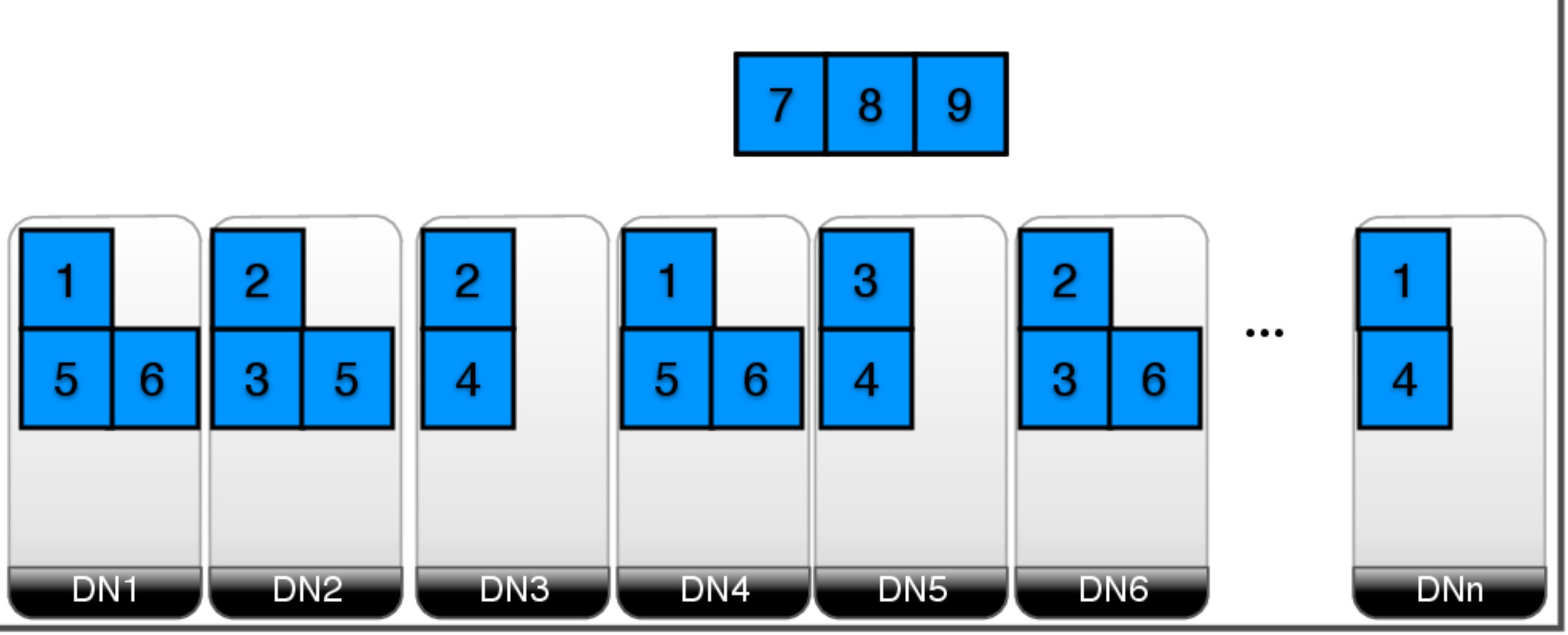
1	2	3
4	5	6
7	8	9

DN1 DN2 DN3 DN4 DN5 DN6 ... DNn

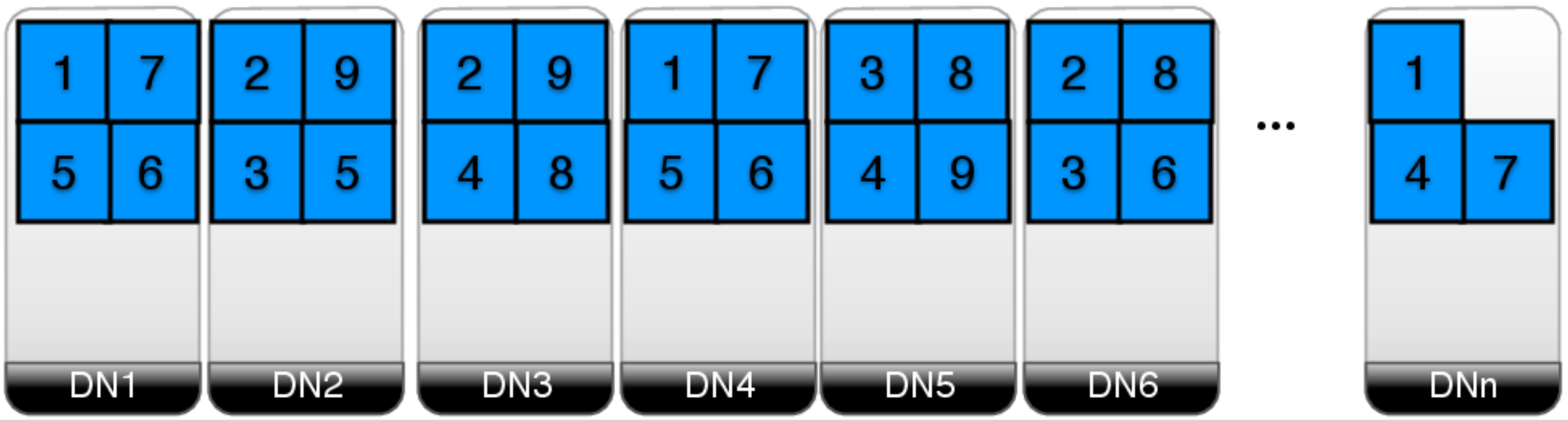
HDFS



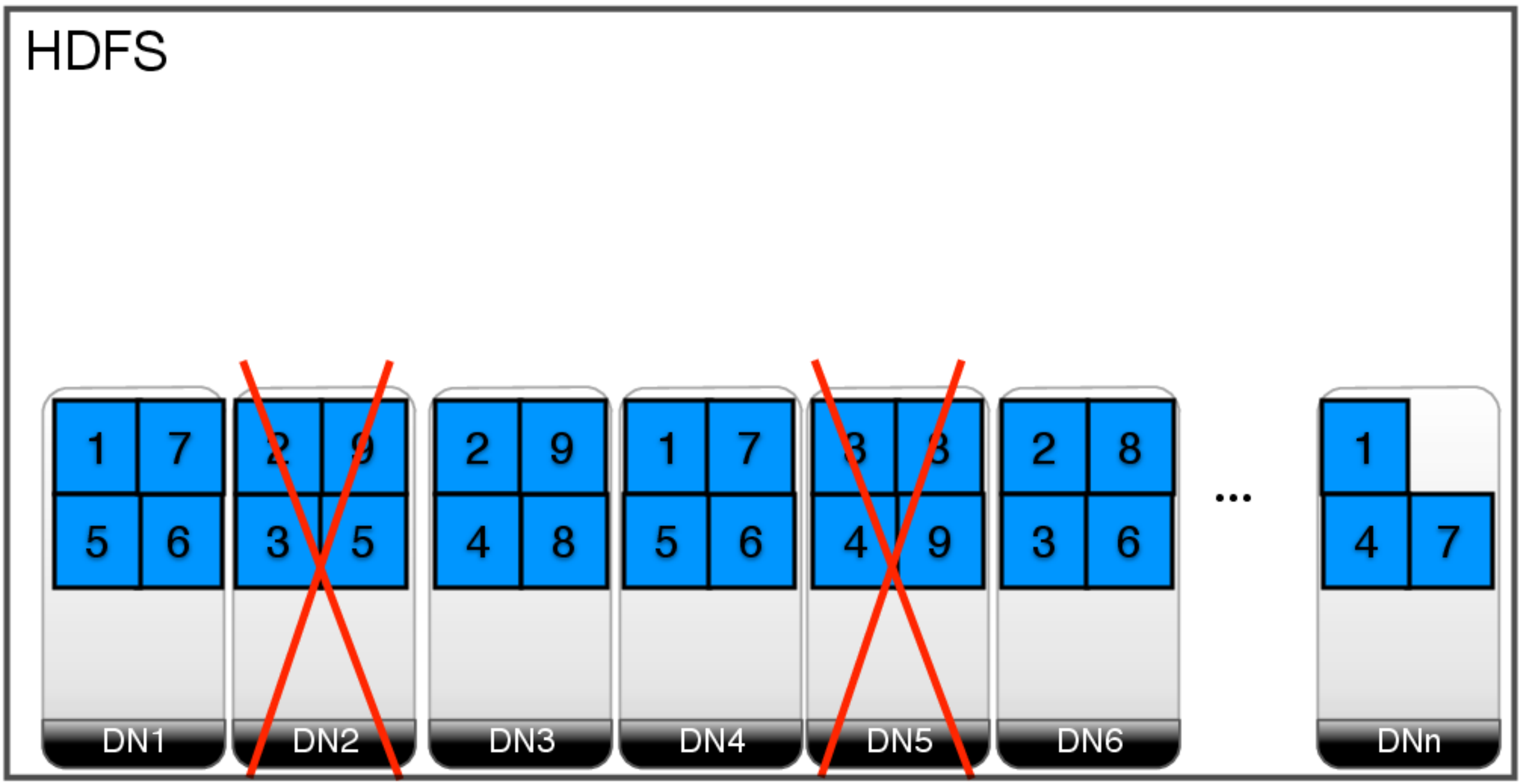
HDFS



HDFS



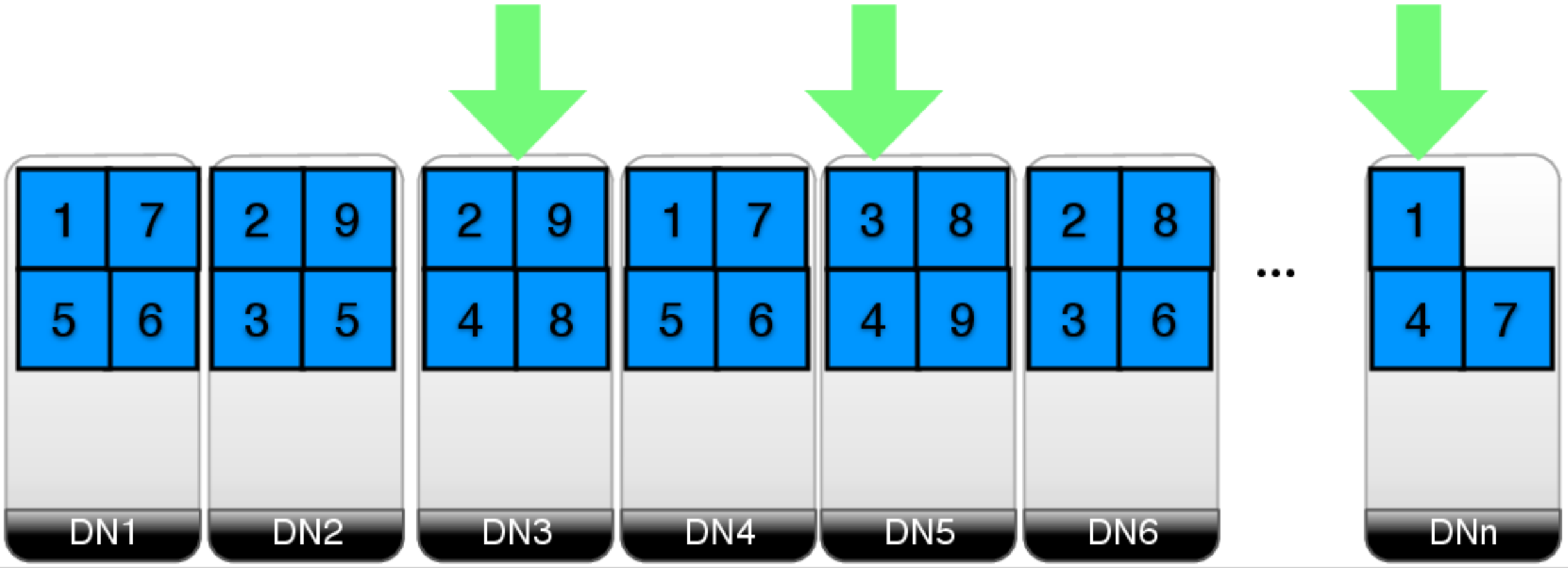
Failover


HDFS
1
7
5
6
2
9
3
5
2
9
4
8
1
7
5
6
3
3
4
9
2
8
3
6
...
1
4
7
DN1
DN2
DN3
DN4
DN5
DN6
DNn

Load Balancing

I would like to have block 4!

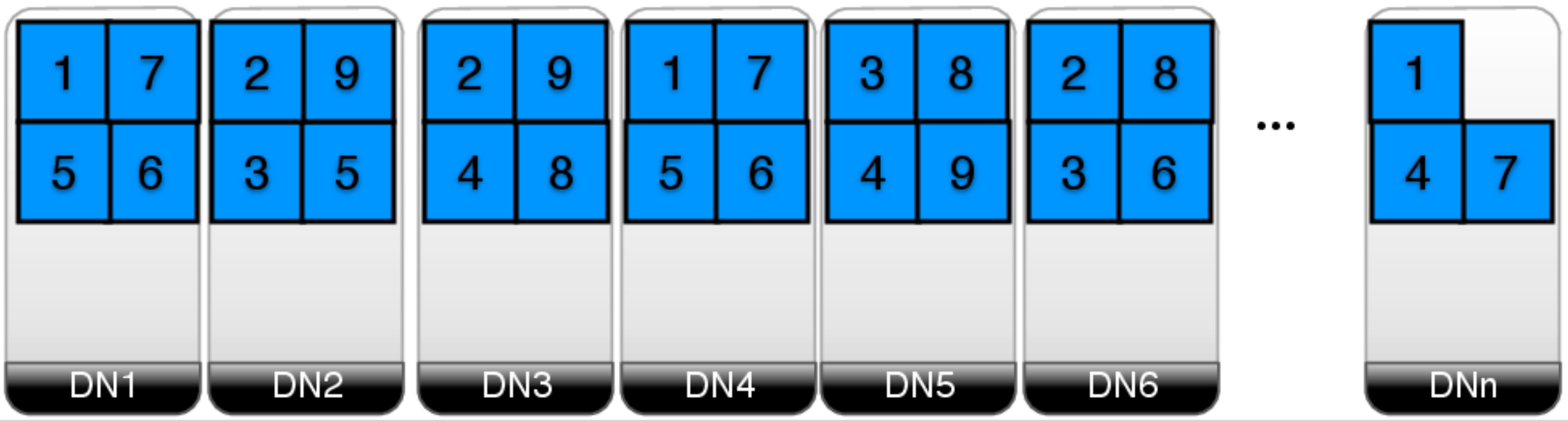
HDFS



Map Phase

MapReduce map((docID, document)) -> set of (term, docID)

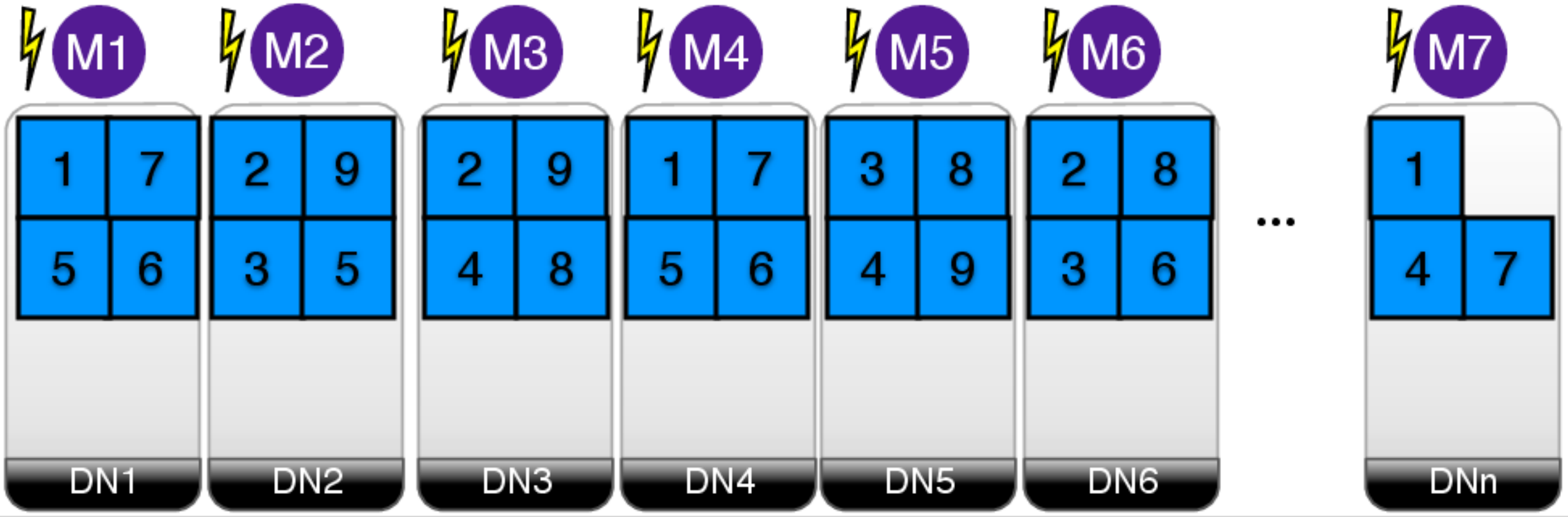
HDFS



Map Phase

MapReduce map((docID, document)) -> set of (term, docID)

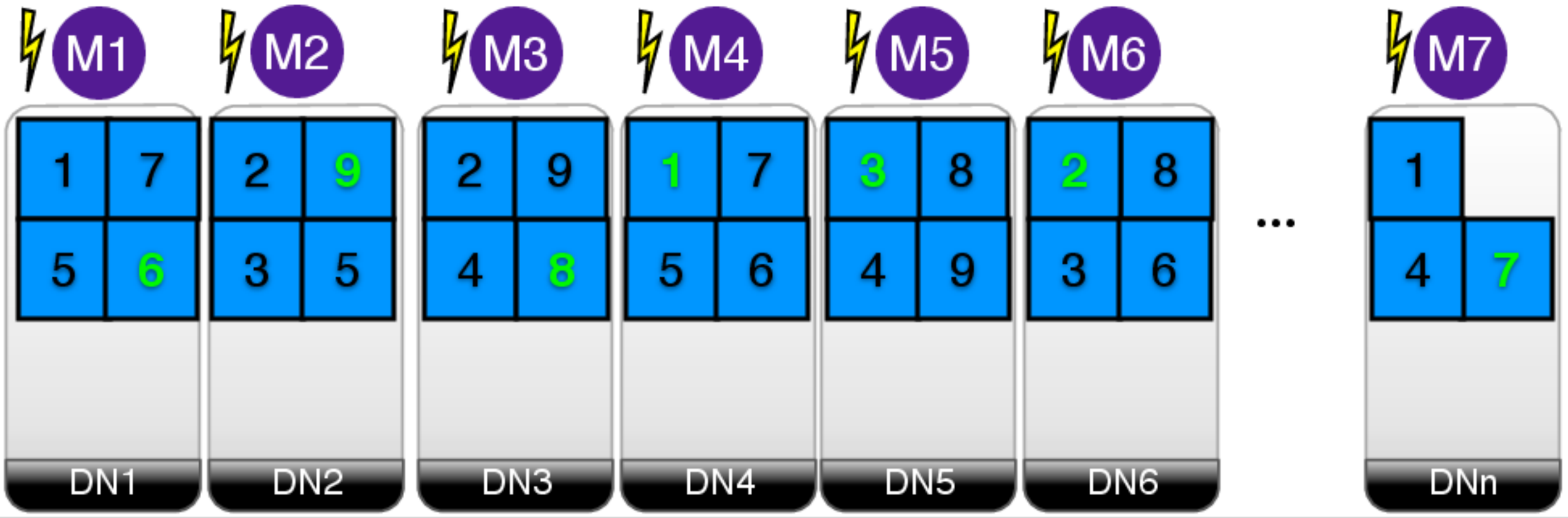
HDFS



Map Phase

MapReduce map((docID, document)) -> set of (term, docID)

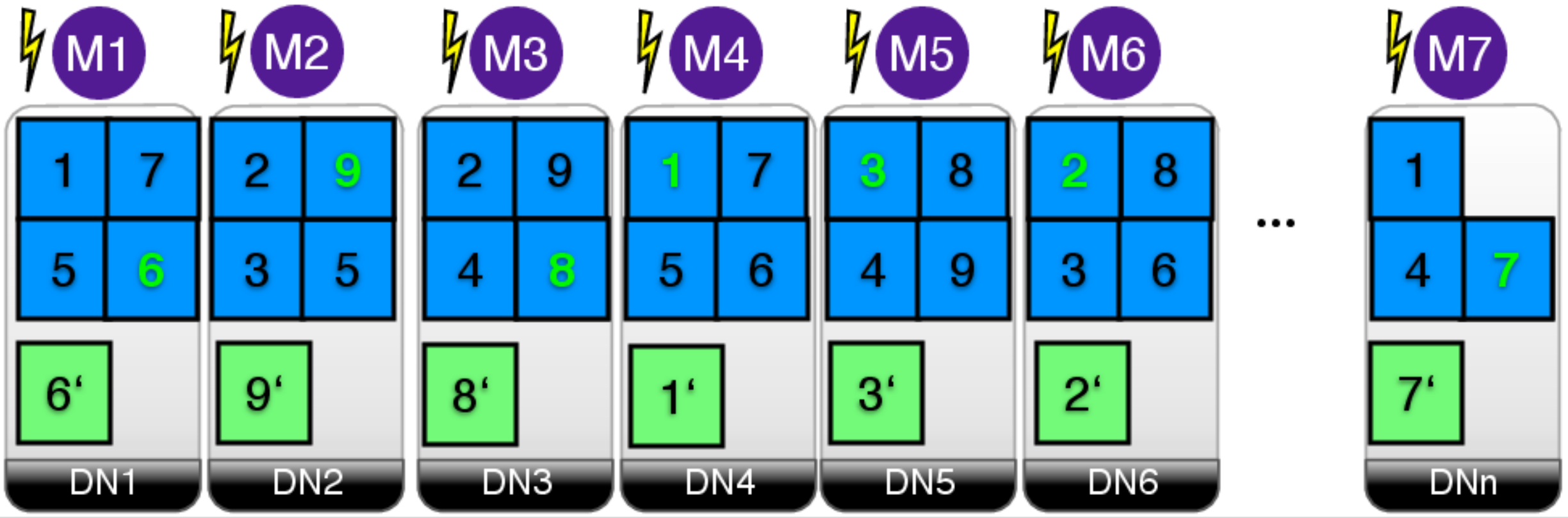
HDFS



Map Phase

MapReduce map((docID, document)) -> set of (term, docID)

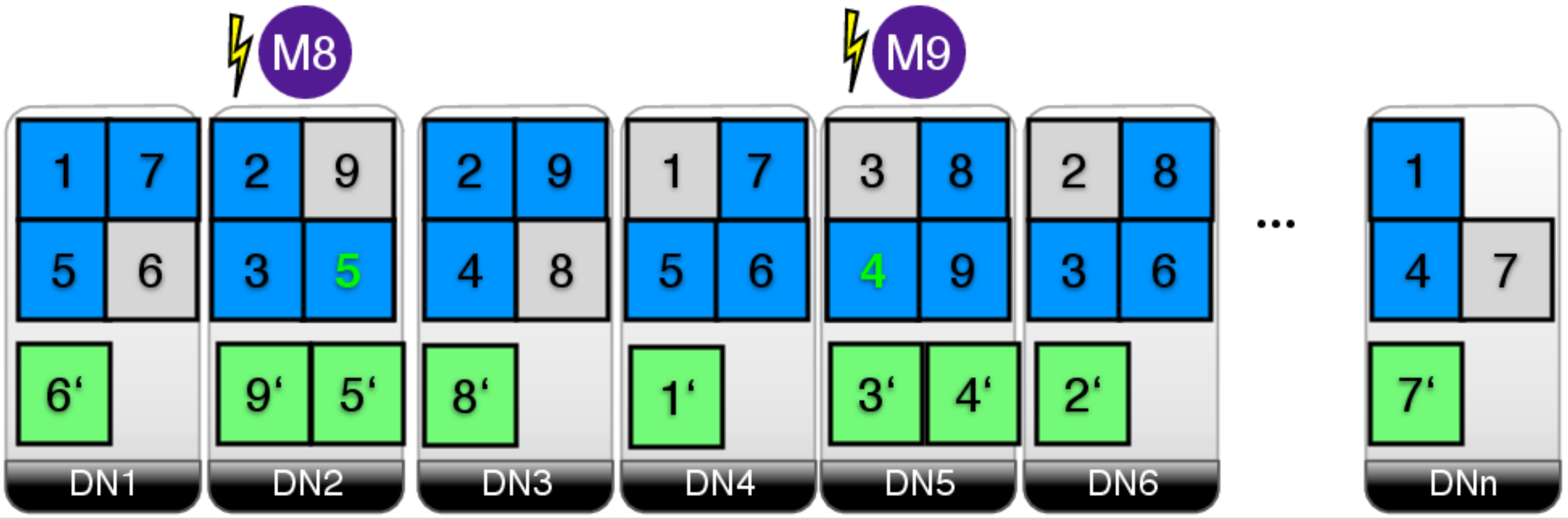
HDFS



Map Phase

MapReduce map((docID, document)) -> set of (term, docID)

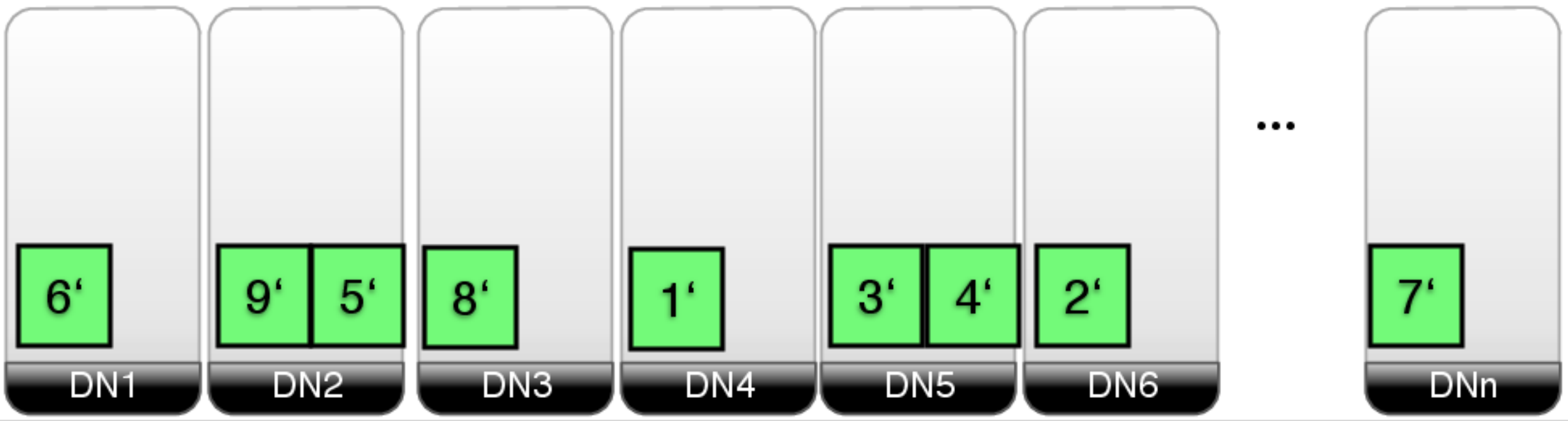
HDFS



Shuffle Phase

MapReduce group by term

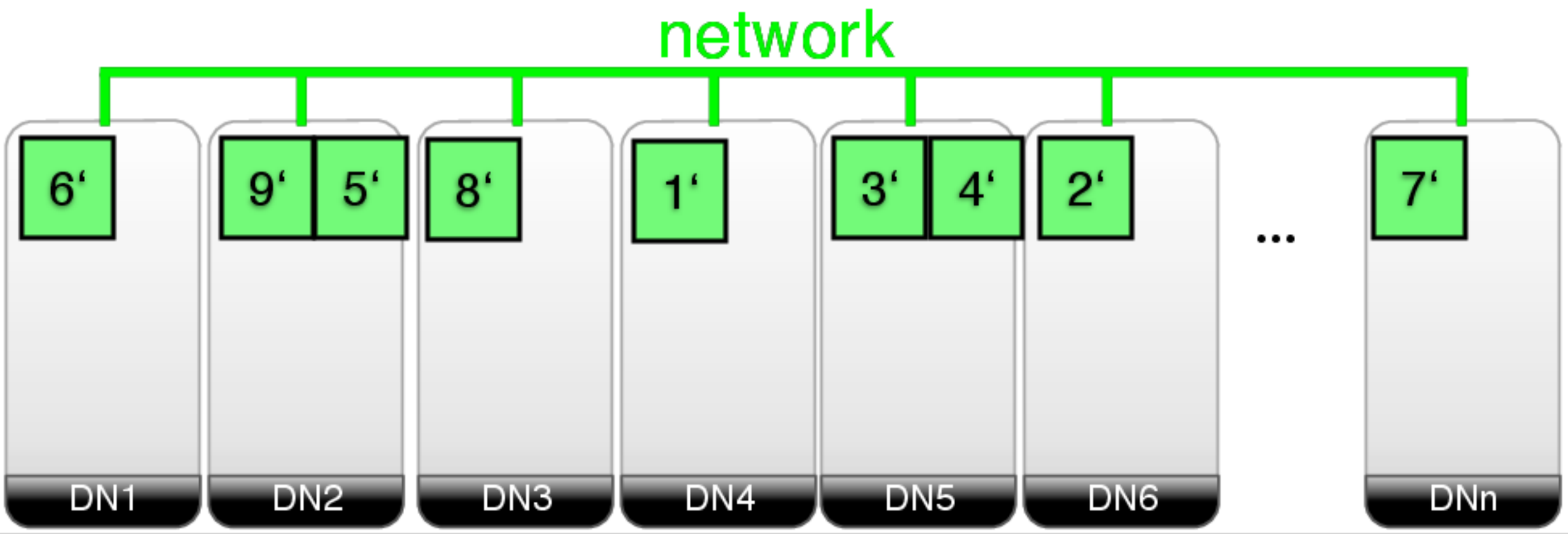
HDFS



Shuffle Phase

MapReduce group by term

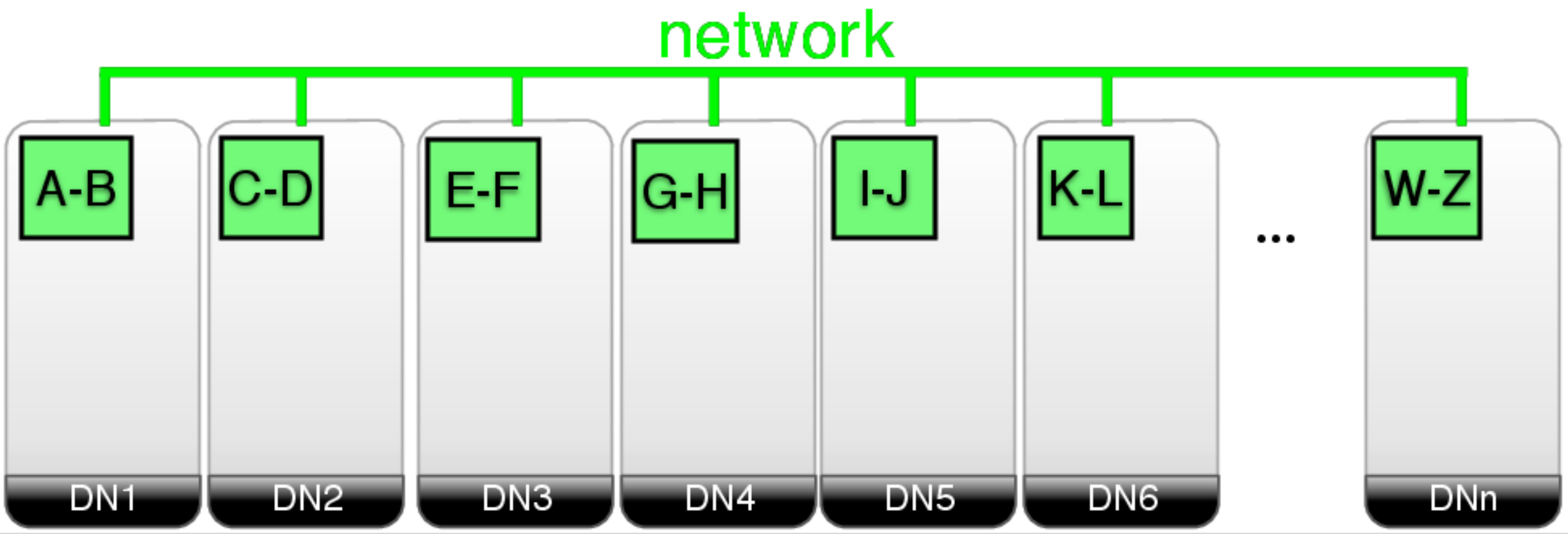
HDFS



Shuffle Phase

MapReduce group by term

HDFS

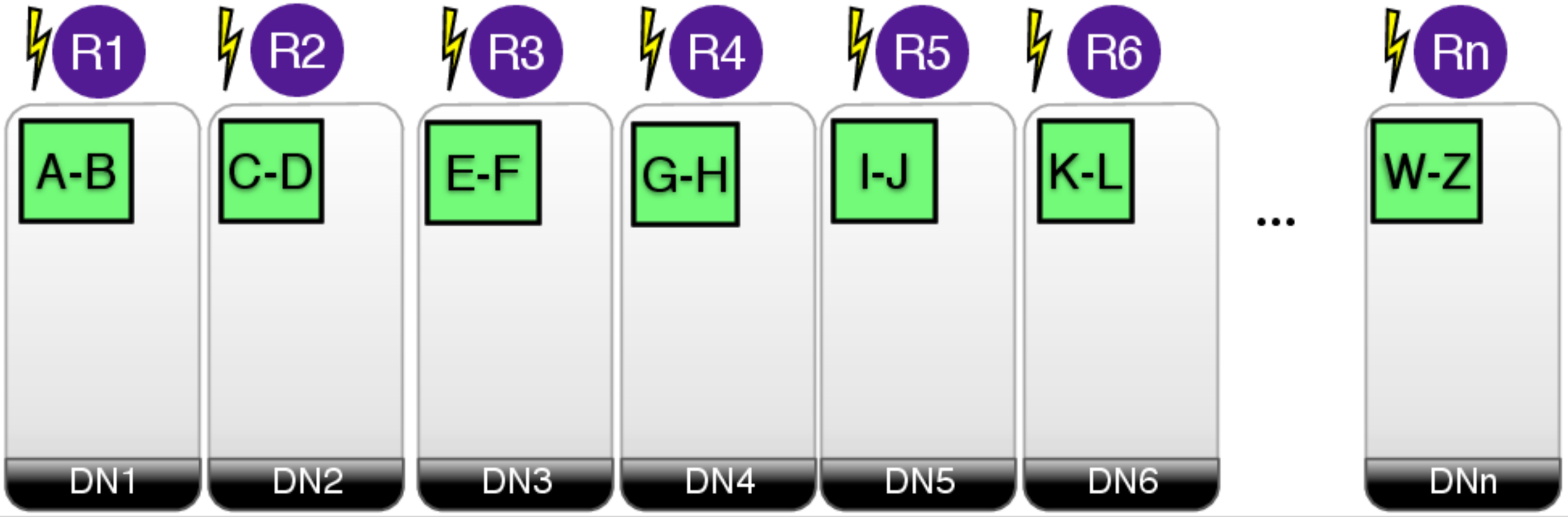


Reduce Phase

reduce(term, set of docID) -> set of (term, (posting list of docID, count))

MapReduce

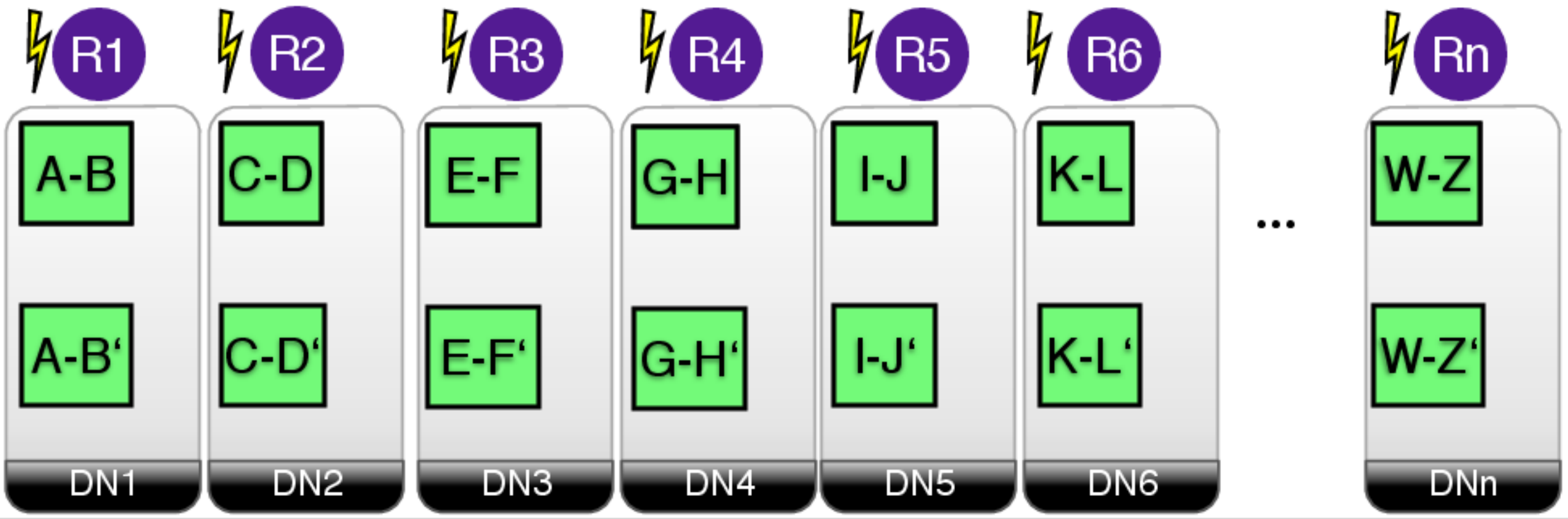
HDFS



Reduce Phase

MapReduce	reduce(term, set of docID) -> set of (term, (posting list of docID, count))

HDFS



Copyrights and Credits

© iStock.com:

hidesy