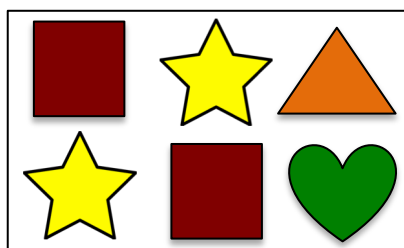
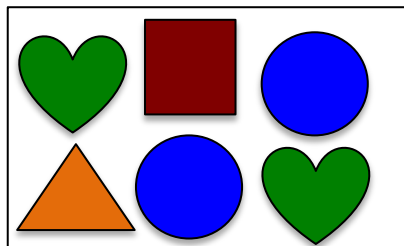
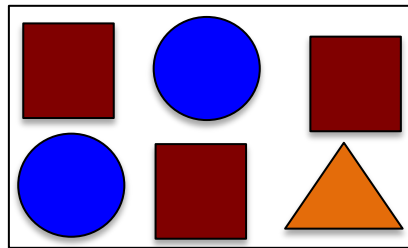
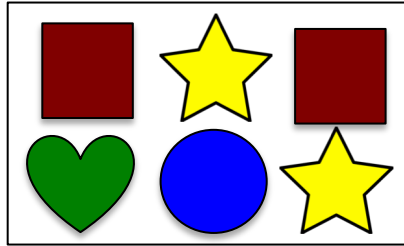





STEP 0 – STORE TO HDFS





Assume 4 data
partitions.



1 - MAP





( , 2)
( , 2)
( , 1)
( , 1)






( , 3)
( , 2)
( , 1)







( , 1)
( , 2)
( , 2)
( , 1)

( , 2)
( , 2)
( , 1)
( , 1)

2 – SHUFFLE and SORT

( , 2)
( , 3)
( , 1)
( , 2)

( , 1)
( , 2)
( , 2)
( , 2)
( , 2)

( , 1)
( , 2)
( , 1)
( , 1)
( , 1)
( , 1)

3 - REDUCE

( , 8)
( , 4)
( , 5)
( , 4)
( , 3)

Note: In the map step I consolidated values and this is not correct according to the slides. I did this to save space and avoid visual clutter.

In reality it should be:
sqr, sqr -> (sqr,1), (sqr,1)
cr, cr, cr -> (cr,1),(cr,1),(cr,1)
And so on...