Introduction

Hadoop, HDFS, MapReduce

Geek Night, TW Gurgaon

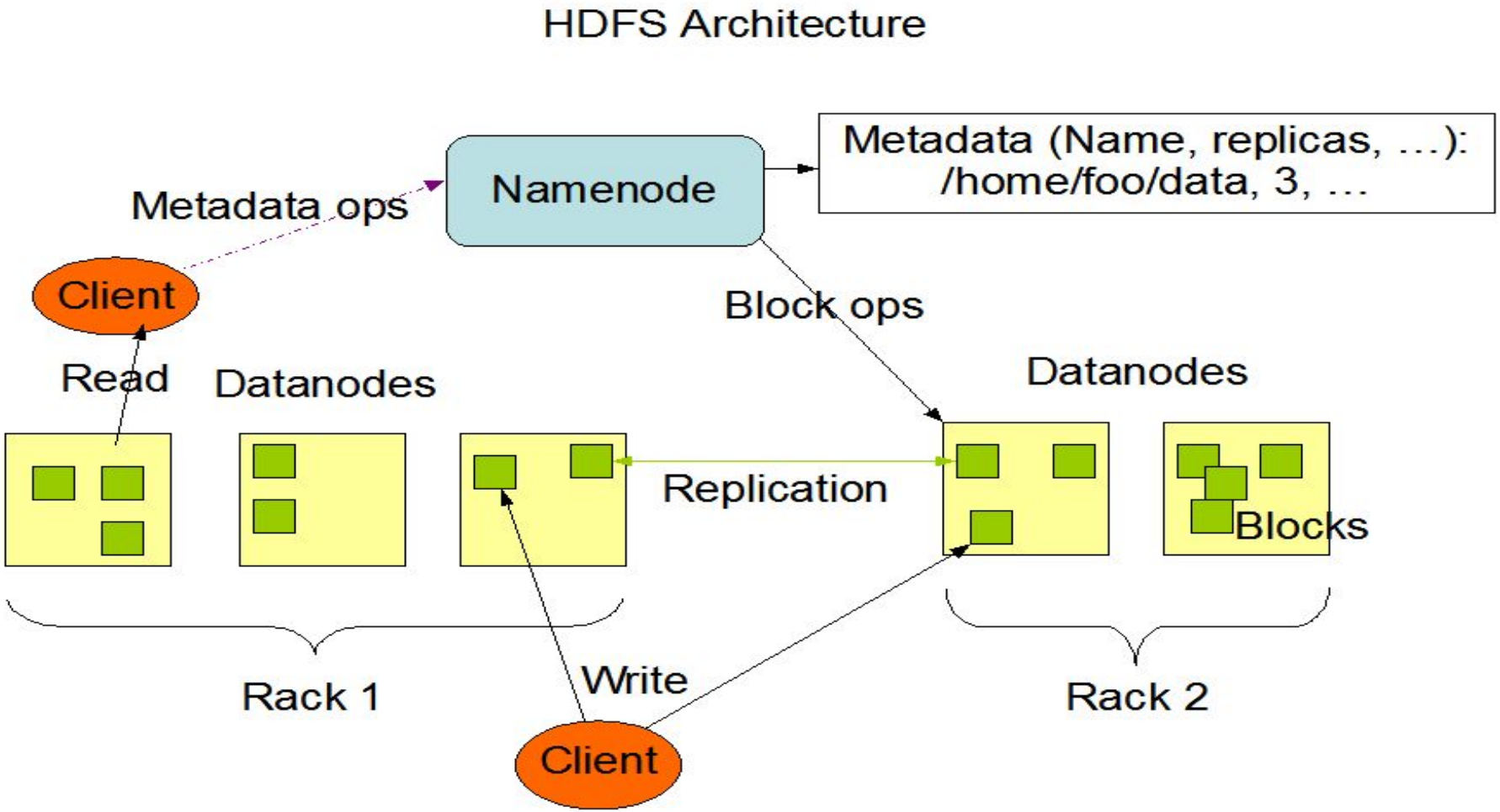
Hadoop

- Open source framework
- Used for storage and large scale processing of data-sets on clusters of commodity hardware
- Mainly consists of the following two modules:

  - HDFS (Distributed Storage)
  - Map Reduce (Analysis/Processing)

HDFS

- Distributed File System based on Google File System (GFS)
- Well suited for commodity hardware
- Built around the idea “Write Once, Read Multiple Times”
- Reliability through replication

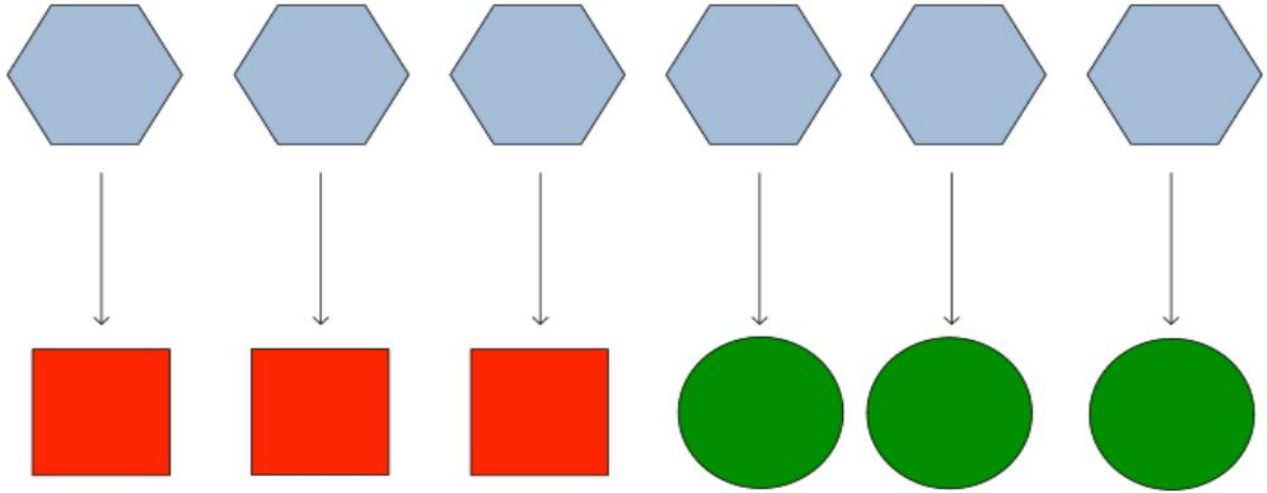

HDFS Architecture
Namenode
Metadata (Name, replicas, …):
/home/foo/data, 3, …
Metadata ops
Client
Block ops
Read
Datanodes
Datanodes
Replication
Blocks
Rack 1
Write
Rack 2
Client

Map-Reduce

- Massively Parallel programming paradigm
- Fault tolerant
- Designed to run on clusters of commodity hardware
- Provides high level of abstraction to developers

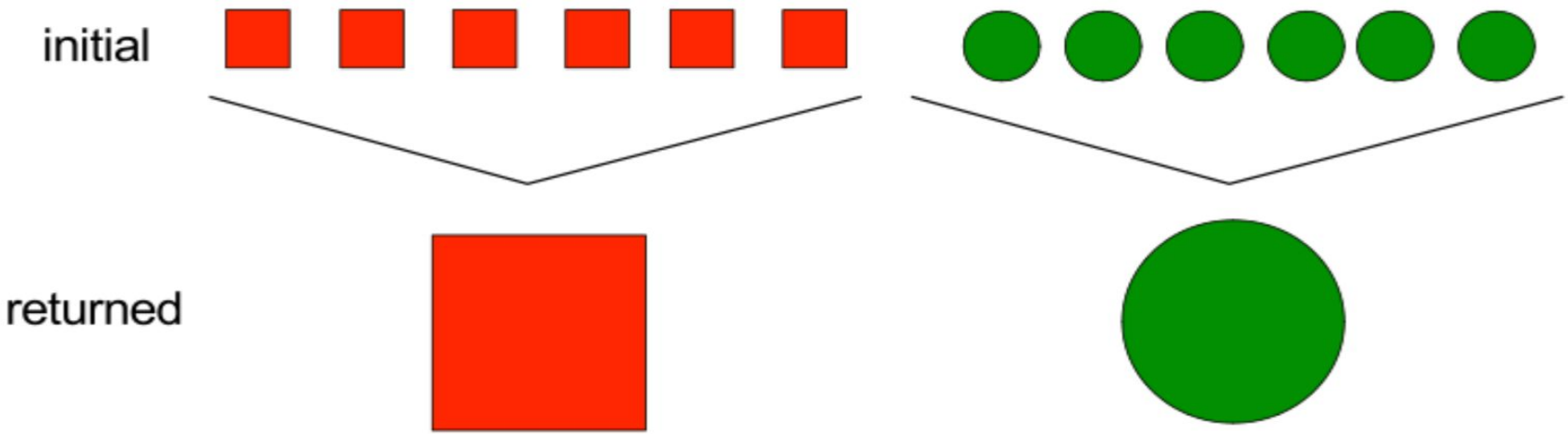
Map

map(K1, V1) -> (K2, V2)

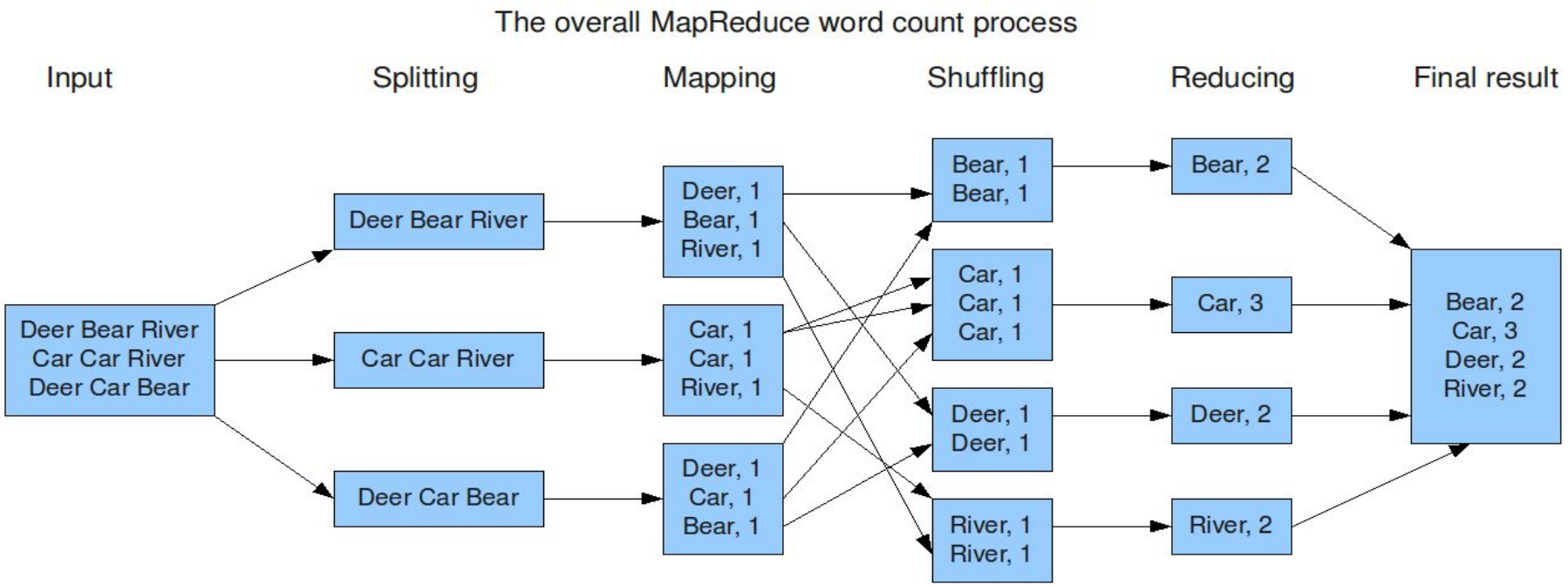


Reduce

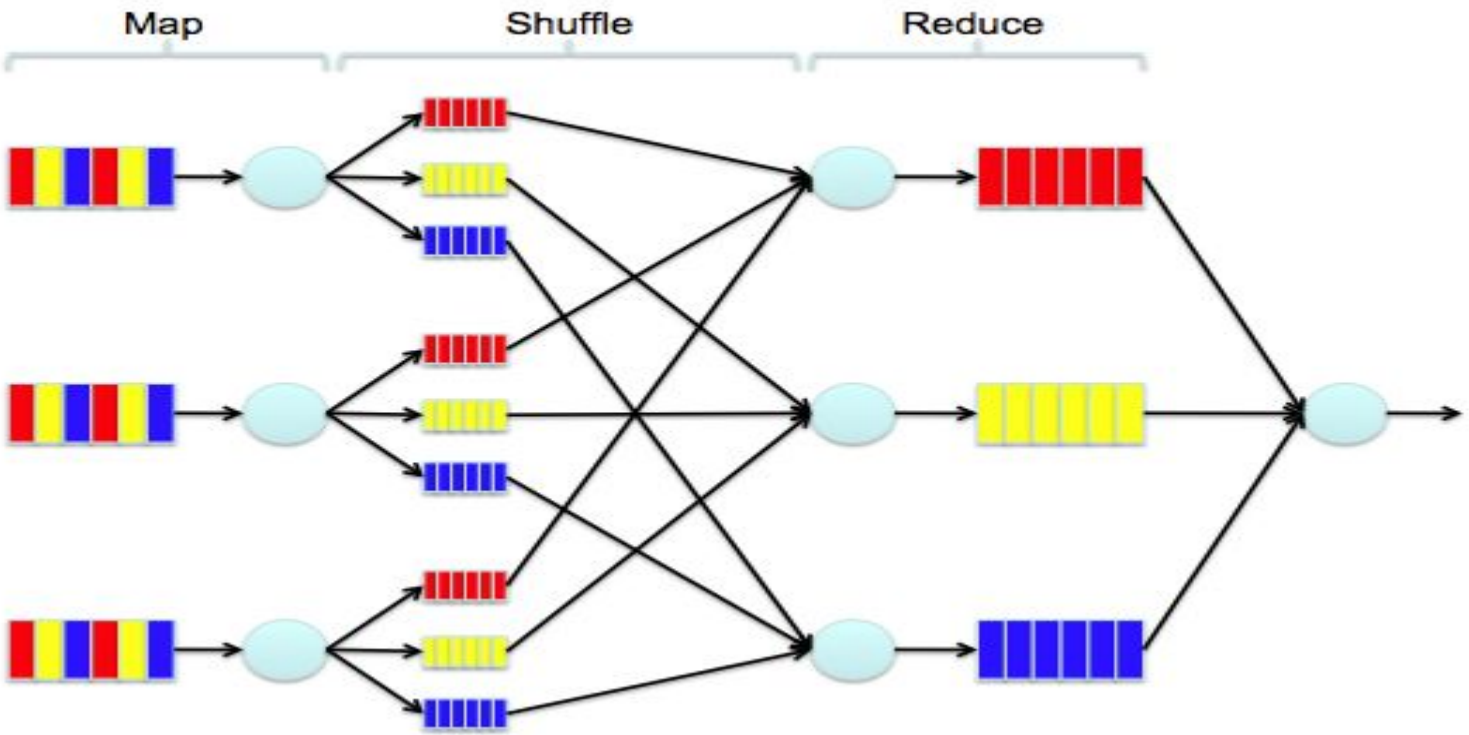
reduce(K2, list(V2) -> list(K3, V3)



The Famous WordCount



Generalized Flow




Let’s Code It Now! :)